Journal of Machine Learning Research 23 (2022) 1-69 Submitted 4/20; Revised 9/21; Published 2/22

Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting

Maxime Vono* MAXIME.VONO@HUAWEI.COM
Lagrange Mathematics and Computing Research Center, Huawei
75007 Paris, France

Daniel Paulin* PAULINDANI@GMAIL.COM
School of Mathematics
University of Edinburgh, United Kingdom

Arnaud Doucet DOUCET@STATS.OX.AC.UK
Department of Statistics
University of Oxford, United Kingdom

Editor: Arnak Dalalyan

Abstract

Performing exact Bayesian inference for complex models is computationally intractable. Markov chain Monte Carlo (MCMC) algorithms can provide reliable approximations of the posterior distribution but are expensive for large data sets and high-dimensional models. A standard approach to mitigate this complexity consists in using subsampling techniques or distributing the data across a cluster. However, these approaches are typically unreliable in high-dimensional scenarios. We focus here on a recent alternative class of MCMC schemes exploiting a splitting strategy akin to the one used by the celebrated alternating direction method of multipliers (ADMM) optimization algorithm. These methods appear to provide empirically state-of-the-art performance but their theoretical behavior in high dimension is currently unknown. In this paper, we propose a detailed theoretical study of one of these algorithms known as the split Gibbs sampler. Under regularity conditions, we establish explicit convergence rates for this scheme using Ricci curvature and coupling ideas. We support our theory with numerical illustrations.

Keywords: ADMM, approximate Bayesian inference, convergence rates, Markov chain Monte Carlo, splitting

1. Introduction

We are interested in performing Bayesian inference for large data sets and potentially high-dimensional models. For complex models, the posterior distribution is intractable and needs to be approximated. To this end, many Markov chain Monte Carlo (MCMC) schemes have been proposed over the past five years; see for instance Bardenet et al. (2017) for a recent overview.

These methods can be loosely speaking divided into two groups: subsampling-based techniques and divide-and-conquer approaches. Subsampling-based approaches are MCMC techniques that only require accessing a subsample of the observations at each iteration:

∗. Both authors contributed equally.

these include the popular stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011; Dubey et al., 2016; Brosse et al., 2018; Chatterji et al., 2018; Baker et al., 2019), subsampling versions of the Metropolis–Hastings algorithm (Bardenet et al., 2014; Korattikara et al., 2014; Bardenet et al., 2017; Quiroz et al., 2019; Cornish et al., 2019) and methods based on piecewise-deterministic MCMC schemes (Bouchard-Côté et al., 2018; Bierkens et al., 2019). However, all the subsampling methods accessing $\mathcal{O}(1)$ data points at each iteration only provide reliable posterior approximations if they rely on some control variate ideas which require estimating the mode of the posterior and this posterior to be concentrated (Welling and Teh, 2011; Dubey et al., 2016; Bardenet et al., 2017; Brosse et al., 2018; Chatterji et al., 2018; Baker et al., 2019; Cornish et al., 2019). Practically, as pointed out in Bardenet et al. (2017); Cornish et al. (2019), this means that such methods are of limited practical interest as they only work well in scenarios where the Bernstein-von Mises approximation of the target is excellent. Divide-and-conquer methods are techniques which consider the common scenario where the data are distributed across a cluster. These schemes run independent MCMC chains to estimate "local" posteriors on each node of the cluster and then recombine these "local" posteriors to obtain an approximation of the full posterior (Wang and Dunson, 2013; Neiswanger et al., 2014; Minsker et al., 2014; Wang et al., 2015; Scott et al., 2016; Scott, 2017; Hasenclever et al., 2017). However, these methods often use parametric or kernel density approximations of the local posteriors so as to combine them. This can be unreliable in high-dimensional scenarios; see Bardenet et al. (2017) and Rendell et al. (2021) for a detailed discussion.

An alternative approach to perform MCMC, amenable to a distributed implementation, has been recently introduced independently in Vono et al. (2019) and Rendell et al. (2021); see also Dai Pra et al. (2012); Chowdhury and Jermaine (2018) and Barbos et al. (2017) for earlier related ideas. It is inspired by the well-known variable splitting technique used in optimization, for instance by quadratic penalty approaches or the alternating direction method of multipliers (ADMM), see Boyd et al. (2011). In the sampling context, this corresponds to defining an artificial hierarchical Bayesian model where the parameter of interest is becoming a "master" parameter which is artificially replicated as many times as one "splits" the target distribution. In this context, we can then develop MCMC schemes which alternate sampling the node parameters given the master parameter then the master parameter given the node parameters. Experimentally, these methods appear promising but it is yet unclear how such schemes behave in high-dimensional scenarios. This paper aims at studying theoretically one of these samplers called split Gibbs sampler (SGS).

Contributions. Our contributions are as follows.

- We present non-asymptotic bounds on the total variation (TV) and 1-Wasserstein distances between the original posterior distribution and the distribution targeted by this class of MCMC schemes. This allows us to quantify the "bias" introduced by these methods and significantly sharpens and complements previous results in Vono et al. (2021a).

- Using Ricci curvature and coupling techniques, we establish explicit dimension-free convergence rates for SGS. Combining our bounds on the bias and convergence rates, we provide mixing time bounds with explicit dependencies with respect to (w.r.t.) the dimension of the problem, its associated condition number and the prescribed preci-

sion. In both 1-Wasserstein and TV distances, we show that our complexity results are competitive with those recently derived for MCMC schemes based on Langevin or Hamiltonian dynamics.

- We illustrate these theoretical results on several applications, demonstrating the benefits of SGS over state-of-the-art MCMC approaches.

Notations and conventions. We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. The total variation norm between two probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{f \in \mathbb{M}(\mathbb{R}^d), \|f\|_\infty \leq 1} \left| \int_{\boldsymbol{\theta} \in \mathbb{R}^d} f(\boldsymbol{\theta}) \mathrm{d}\mu(\boldsymbol{\theta}) - \int_{\boldsymbol{\theta} \in \mathbb{R}^d} f(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta}) \right|,$$

where $\mathbb{M}(\mathbb{R}^d)$ denotes the set of all Borel measurable functions f on $\mathbb{R}^d$ and $\|f\|_\infty = \sup_{\boldsymbol{\theta} \in \mathbb{R}^d} |f(\boldsymbol{\theta})|$. Let μ, ν be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Define the Kullback-Leibler (KL) divergence of μ from ν by

$$D_{\mathrm{KL}}(\mu || \nu) = \begin{cases} \int_{\mathbb{R}^d} \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\boldsymbol{\theta})\right) \mathrm{d}\nu(\boldsymbol{\theta}), & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise.} \end{cases}$$

For $1 \leq p < \infty$, and a metric $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the Wasserstein distance of order p between two probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is defined by

$$W_p^w(\mu, \nu) = \left(\inf_{\pi \in \mathcal{U}(\mu,\nu)} \int_{\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d} w(\boldsymbol{\theta}, \boldsymbol{\theta}')^p \mathrm{d}\pi(\boldsymbol{\theta}, \boldsymbol{\theta}') \right)^{1/p},$$

where $\mathcal{U}(\mu, \nu)$ is the set of all probability measures which admit μ and ν as marginals. For $p = \infty$, the Wasserstein distance of order ∞ is defined as

$$W_\infty^w(\mu, \nu) = \inf_{\pi \in \mathcal{U}(\mu,\nu), (X,Y) \sim \pi} \operatorname{ess\,sup} w(X, Y) \ .$$

In the case when w is the Euclidean metric, we will denote these by $W_p(\mu, \nu)$. For the sake of simplicity, with little abuse, we shall use the same notations for a probability distribution and its associated probability density function. For a Markov chain with transition kernel $\mathbf{P}$ on $\mathbb{R}^d$ and invariant distribution π, we define the ϵ-mixing time associated to a statistical distance D, precision $\epsilon > 0$ and initial distribution ν, by

$$t_{\mathrm{mix}}(\epsilon; \nu) = \min \left\{ t \geq 0 \ \middle| \ D(\nu \mathbf{P}^t, \pi) \leq \epsilon \right\},$$

which stands for the minimum number of steps of the Markov chain such that its distribution is at most at an ϵ D-distance from the invariant distribution π. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$. For $n \geq 1$, we refer to the set of integers between 1 and n with the notation $[n]$. The d-multidimensional Gaussian probability distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The parabolic cylinder special function is defined, for all $d > 0$ and $z \in \mathbb{R}$, by $D_{-d}(z) = \exp(-z^2/4)\Gamma(d)^{-1} \int_0^{+\infty} e^{-xz - x^2/2} x^{d-1} \mathrm{d}x$, where $\Gamma(\cdot)$ denotes the Gamma function. For $0 \leq i < j$, we use the notation $\mathbf{u}_{i:j}$ to refer to the vector $[\mathbf{u}_i^\top, \mathbf{u}_{i+1}^\top, \ldots, \mathbf{u}_j^\top]^\top$ built by stacking $j - i + 1$ vectors $(\mathbf{u}_k; k \in \{i, i+1, \ldots, j\})$. For $f : \mathbb{R}^d \to \mathbb{R}$ and any $\boldsymbol{\theta} \in \mathbb{R}^d$, we use the notations $f(\boldsymbol{\theta})_- = -\min(f(\boldsymbol{\theta}), 0)$ and $f(\boldsymbol{\theta})_+ = \max(f(\boldsymbol{\theta}), 0)$.

2. Background and Problem Formulation

This section sets up the simulation problem considered in this paper and briefly reviews the approximate Bayesian approach proposed by Vono et al. (2019) and Rendell et al. (2021).

2.1 Bayesian Model

We consider the situation where one is interested in carrying out Bayesian inference about a parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ based on observed data $\mathrm{D} = \{\mathbf{x}_j, y_j\}_{j=1}^n$, where for any $j \in [n]$, $\mathbf{x}_j$ are covariates (also called features) associated to observation y_j. We assume that the number of observations n is large so that the data set D is partitioned into $S \in [n]$ subsets $\{\mathrm{D}_s\}_{s=1}^S$, called *shards*, such that $\sqcup_{s=1}^S \mathrm{D}_s = \mathrm{D}$. Under this framework, the posterior distribution of interest is assumed to admit a density w.r.t. the Lebesgue measure of the form

$$\pi(\boldsymbol{\theta} \mid \mathrm{D}) \propto p(\boldsymbol{\theta}) \prod_{s=1}^{S} \pi_s(\mathrm{D}_s \mid \boldsymbol{\theta}) \ , \tag{1}$$

where $\{\pi_s(\mathrm{D}_s \mid \boldsymbol{\theta})\}_{s=1}^S$ are likelihood functions associated to $\{\mathrm{D}_s\}_{s=1}^S$ and $p(\boldsymbol{\theta})$ is the prior density for $\boldsymbol{\theta}$. Contrary to the majority of divide-and-conquer MCMC approaches, we do not assume that the prior factorizes across shards, that is $p(\boldsymbol{\theta}) \propto \prod_{s=1}^S p_s(\boldsymbol{\theta})$. In the sequel, it will be convenient to characterize the posterior distribution via potential functions. To this end, we assume that the posterior density defined in (1) can be re-written as $\pi(\boldsymbol{\theta} \mid \mathrm{D}) \propto \exp(-U(\boldsymbol{\theta}))$ where

$$U(\boldsymbol{\theta}) = \sum_{i=1}^{b} U_i(\mathbf{A}_i \boldsymbol{\theta}) \ , \tag{2}$$

for $b \in \mathbb{N} \setminus \{0\}$, some matrices $\mathbf{A}_i \in \mathbb{R}^{d_i \times d}$ and potential functions $U_i : \mathbb{R}^{d_i} \to \mathbb{R}$ with $i \in [b]$. This definition of the posterior encompasses two main scenarios that are ubiquitous in Bayesian machine learning and illustrated in Examples 1 and 2. More precisely, if $p(\boldsymbol{\theta}) \propto \prod_{s=1}^S p_s(\boldsymbol{\theta})$, then $b = S$ and for any $i \in [b]$, $U_i(\mathbf{A}_i\boldsymbol{\theta}) = -\log p_i(\boldsymbol{\theta}) - \log \pi_i(\mathrm{D}_i \mid \boldsymbol{\theta})$. Conversely, if $p(\boldsymbol{\theta})$ does not factorize across shards, one can set $b = S + 1$ by assigning, for $s \in [S]$, one potential U_s to each likelihood contribution $\pi_s(\mathrm{D}_s \mid \boldsymbol{\theta})$, and one potential U_{S+1} to the prior $p(\boldsymbol{\theta})$. In both cases, for any $i \in [b]$, the potential U_i is assumed to be dependent on the subset D_i of the observations; potentially $\mathrm{D}_i = \{\emptyset\}$ if U_i refers to the prior. To simplify notation, this dependence is notationally omitted and we will denote by $\pi(\boldsymbol{\theta})$ the posterior distribution in the rest of the paper. We give hereafter two illustrative standard statistical machine learning examples that fit into the considered Bayesian framework.

Example 1 *Bayesian ridge linear regression. We consider the model defined by*

$$\begin{aligned} &y_j \sim \mathcal{N}(\mathbf{x}_j^\top \boldsymbol{\theta}, \sigma^2) \ , \quad \forall j \in [n] \ , \\ &\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}_d, \tau \mathbf{I}_d\right) \ , \end{aligned}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ are the unknown regression parameters and $\tau > 0$ is a fixed regularization parameter. In this case, the posterior density writes

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tau}\left\|\boldsymbol{\theta}\right\|^2\right) \prod_{j=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_j - \mathbf{x}_j^\top \boldsymbol{\theta})^2\right) \ .$$

By dividing the data set $\mathrm{D} = \{\mathbf{x}_j, y_j\}_{j=1}^n$ *into* S *shards* $\{\mathrm{D}_s\}_{s=1}^b$*, the posterior density can be re-written as in* (1)*, that is*

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\tau}\left\|\boldsymbol{\theta}\right\|^2\right) \prod_{s=1}^{S} \exp\left(-\frac{1}{2\sigma^2} \sum_{\{\mathbf{x}_j, y_j\} \in \mathrm{D}_s} (y_j - \mathbf{x}_j^\top \boldsymbol{\theta})^2\right) \ .$$

Under this factorization, one can characterize $\pi(\boldsymbol{\theta})$ *via* (2) *by setting* $b = S + 1$ *with the choices* $\mathbf{A}_{S+1} = \mathbf{I}_d$*,* $U_{S+1} = \left\|\boldsymbol{\theta}\right\|^2/(2\tau)$*, and for any* $s \in [S]$*,* $\mathbf{A}_s = \mathbf{I}_d$*,* $U_s(\boldsymbol{\theta}) = \sum_{\{\mathbf{x}_j, y_j\} \in \mathrm{D}_s} (y_j - \mathbf{x}_j^\top \boldsymbol{\theta})^2/(2\sigma^2)$*. In this case, note that* $D_{S+1} = \{\emptyset\}$*. Robust linear regression also falls into this framework. In this case, for any* $j \in [n]$*,* y_j *is distributed according to Student's t-distribution.*

Example 2 *Bayesian logistic regression with Zellner prior. Consider the model defined by*

$$y_j \sim \text{Bernoulli}\left(\sigma\left(\mathbf{x}_j^\top \boldsymbol{\theta}\right)\right) \ , \ \forall j \in [n] \ , \tag{3}$$

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}_d, \boldsymbol{\Sigma}\right) \ , \tag{4}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ *are the unknown regression parameters,* $\sigma(u) = 1/(1 + \mathrm{e}^{-u})$ *is the logistic link, and* $\boldsymbol{\Sigma}^{-1} = \alpha \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top$*, with* $\alpha = 3d/(\pi^2 n)$ *which corresponds to a Zellner prior (Sabanes Bove and Held, 2011; Hanson et al., 2014). In this scenario, the posterior density writes* $\pi \propto \mathrm{e}^{-U}$ *with*

$$U(\boldsymbol{\theta}) = \sum_{j=1}^{n} y_j \mathbf{x}_j^\top \boldsymbol{\theta} + \log\left[1 + \exp\left(-\mathbf{x}_j^\top \boldsymbol{\theta}\right)\right] + \frac{\alpha}{2}\left\|\mathbf{x}_j^\top \boldsymbol{\theta}\right\|^2 \ .$$

This posterior density can be re-written as in (2) *by setting* $b = n$ *and for* $i \in [b]$*,* $d_i = 1$*,* $\mathbf{A}_i = \mathbf{x}_i^\top$ *and* $U_i(u) = y_i u + \log(1 + \mathrm{e}^{-u}) + \alpha u^2/2$*. In this case,* $\mathrm{D}_i = \{\mathbf{x}_i, y_i\}$ *for any* $i \in [b]$*. Similarly to the logistic regression, other Bayesian generalized linear models such as multinomial logistic regression and Poisson regression also fall into this framework, see McCullagh and Nelder (2019) for more examples.*

Sampling from π defined in (2) is challenging because both the number of data n and the dimension d can be large. In addition, the data set D might be distributed over a cluster, which complicates the inference procedure.

2.2 Instrumental Hierarchical Bayesian Model

To address these issues, Vono et al. (2019) and Rendell et al. (2021) introduced an artificial/instrumental Bayesian hierarchical model to ease posterior computation. The idea is to introduce an auxiliary variable $\mathbf{z}_i \in \mathbb{R}^{d_i}$ for some factors $i \in [b]$ such that, under an instrumental prior distribution, these variables are conditionally independent given $\boldsymbol{\theta}$. Depending on the structure of the initial posterior distribution, different instrumental hierarchical models have been considered by the aforementioned authors. In this paper, we will study the instance where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{A}_i\boldsymbol{\theta}, \rho^2 \mathbf{I}_{d_i})$ for some $\rho > 0$. Under this model, the artificial joint posterior distribution $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b}) \propto \exp(-U(\boldsymbol{\theta}, \mathbf{z}_{1:b}))$ admits a potential function

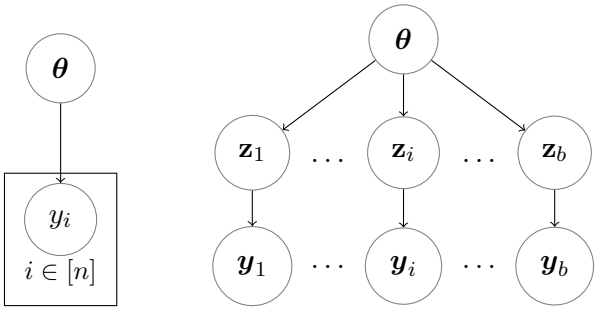

Figure 1: DAGs for (left) the original model (2) and (right) the instrumental model (5). For any $i \in [b]$, the notation $\boldsymbol{y}_i$ refers to the subset $\{y_j \mid y_j \in \mathrm{D}_i\}$. Note that we do not illustrate the dependencies on covariates which can be used to define the matrices $\{\mathbf{A}_i\}_{i=1}^b$ as in Example 2.

U defined by

$$U(\boldsymbol{\theta}, \mathbf{z}_{1:b}) = \sum_{i=1}^{b} U_i(\mathbf{z}_i) + \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2} \ . \tag{5}$$

Figure 1 shows the directed acyclic graph (DAG) associated to this instrumental Bayesian hierarchical model. We could have considered an alternative prior for $\mathbf{z}_i$ as in Dai Pra et al. (2012); Rendell et al. (2021); Vono et al. (2021a) but this choice is motivated by the fact that the corresponding quadratic potential enjoys attractive properties such as smoothness and strong convexity. Conditions ensuring that $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ is a probability density function are detailed in Propositions 3 and 5.

A key property of this artificial posterior distribution is that the resulting *marginal posterior distribution*

$$\pi_\rho(\boldsymbol{\theta}) = \int \Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b}) \mathrm{d}\mathbf{z}_{1:b} \tag{6}$$

converges to the posterior distribution of interest $\pi(\boldsymbol{\theta})$ in total variation norm as $\rho \to 0$. This follows directly from the fact that $\mathcal{N}(\mathbf{z}_i; \mathbf{A}_i\boldsymbol{\theta}, \rho^2\mathbf{I}_{d_i})$ weakly converges towards the Dirac distribution $\delta_{\mathbf{A}_i\boldsymbol{\theta}}(\mathbf{z}_i)$ when $\rho \to 0$ by Scheffé's lemma (Scheffé, 1947).

Another key property of the marginal posterior distribution π_ρ is that it can be expressed in terms of convolutions in an explicit form. Suppose that $\min_{i\in[b]} \inf_{\mathbf{z}_i} U_i(\mathbf{z}_i) > -\infty$ and let

$$U_i^\rho(\mathbf{A}_i\boldsymbol{\theta}) := -\log \int_{\mathbb{R}^{d_i}} \exp\left(-U_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right) \cdot \frac{\mathrm{d}\mathbf{z}_i}{(2\pi\rho^2)^{d_i/2}} \text{ for } i \in [b] \ , \text{ and}$$

$$U^\rho(\boldsymbol{\theta}) := \sum_{i=1}^{b} U_i^\rho(\mathbf{A}_i\boldsymbol{\theta}) \ . \tag{7}$$

Then, Proposition 5 shows that $\pi_\rho(\boldsymbol{\theta}) \propto \exp(-U^\rho(\boldsymbol{\theta}))$ whenever $\exp(-U^\rho(\boldsymbol{\theta}))$ is integrable on $\mathbb{R}^d$.

The instrumental potential (5) slightly generalizes the approach from Vono et al. (2019); Rendell et al. (2021). In Rendell et al. (2021), only the case $\mathbf{A}_i = \mathbf{I}_d$ is considered so that

$\mathbf{z}_i \in \mathbb{R}^{d_i}$ where $d_i = d$. This can be very inefficient. In many applications, we can indeed define auxiliary variables $\mathbf{z}_i$ taking values in $\mathbb{R}^{d_i}$ where $d_i \ll d$. For instance, in the logistic regression example presented in Example 2, we have $d_i = 1$ for $i \in [n]$ while d can be large. Hence, simulation from $\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ is expected to be much cheaper. Efficient sampling from such conditionals is a key ingredient to SGS as described in the next section.

2.3 Split Gibbs Sampler

The main benefit of working with the artificial target distribution $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ defined by (5) instead of $\pi(\boldsymbol{\theta})$ is the fact that, under Π_ρ, the conditional distribution of the auxiliary variables $\mathbf{z}_{1:b}$ given $\boldsymbol{\theta}$ factorizes across $i \in [b]$, that is $\Pi_\rho(\mathbf{z}_{1:b}|\boldsymbol{\theta}) = \prod_{i=1}^b \Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$. Hence these simulation steps can be performed in parallel. This suggests using a Gibbs sampler to sample from $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$. The resulting so-called split Gibbs sampler is described in Algorithm 1. Simple conditions ensuring the ergodicity of SGS are given in Appendix A.1. In the following paragraphs, we detail such conditional sampling problems and discuss the applicability of SGS.

Algorithm 1: Split Gibbs Sampler (SGS)

Input: Potentials $\{U_i\}_{i\in[b]}$, penalty parameter ρ, initialization $\boldsymbol{\theta}^{[0]}$ and nb. of iterations T.

for $t \leftarrow 1$ **to** T **do**
  for $i \leftarrow 1$ **to** b **do**
    $\mathbf{z}_i^{[t]} \sim \Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}^{[t-1]})$ (see Equation 8)
  end
  $\boldsymbol{\theta}^{[t]} \sim \Pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:b}^{[t]})$ (see Equation 11)
end

2.3.1 Sampling the Auxiliary Variables

As emphasized previously, SGS is an attractive sampler since the auxiliary variables $\{\mathbf{z}_i\}_{i=1}^b$ can be sampled in parallel given $\boldsymbol{\theta}$ from the conditional distributions

$$\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}) \propto \exp\left(-U_i(\mathbf{z}_i) - \frac{1}{2\rho^2}\left\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\right\|^2\right) . \tag{8}$$

Additionally each conditional $\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}, \mathbf{y})$ only depends on $\mathbf{y} = \{y_j\}_{j=1}^n$ through the subset of observations D_i, see Figure 1. This is particularly interesting in scenarios where observations $\mathbf{y}$ are distributed over a set of nodes within a cluster such that each node involves the subset $\boldsymbol{y}_i$, see the work by Rendell et al. (2021) for more details. Not only parallel and possibly distributed sampling from (8) is possible but this conditional distribution is far simpler than the original target distribution (2). Indeed, while the latter involves a composite potential function U with matrices $\{\mathbf{A}_i\}_{i=1}^b$ acting on $\boldsymbol{\theta}$, (8) only involves a single potential U_i and an isotropic Gaussian term without any matrix acting on $\mathbf{z}_i$. Hence, sampling from (8) is expected to be easier and cheaper.

In the literature, sampling from this conditional distribution has been performed via two main approaches: exact sampling and Metropolis-Hastings schemes. For example,

the authors in Vono et al. (2019) considered a linear Gaussian inverse problem where (8) was a Gaussian distribution. Apart from the Gaussian case, exact and efficient sampling is for instance possible when considering generalized (non-)linear models and Gaussian prior distributions for $\boldsymbol{\theta}$. Similarly to the Bayesian logistic regression example presented in Example 2, one can indeed assign one potential per univariate observation y_i leading to univariate potentials $\{U_i\}_{i=1}^n$. Hence, in this case one can sample from $\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta}, y_i)$ for $i \in [n]$ by using adaptive rejection sampling (Gilks and Wild, 1992; Martino and Míguez, 2011).

When exact sampling was not possible in practice, the authors in Rendell et al. (2021) considered a Metropolis-Hastings scheme to sample from $\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$. This defines a Metropolis-within-SGS scheme which can be shown to admit Π_ρ as stationary distribution under mild assumptions.

In this paper, we are interested in providing explicit and non-asymptotic theoretical convergence guarantees for Algorithm 1. Since no explicit convergence result exists for special instances of Algorithm 1, we choose to focus on the simplest scenario where exact sampling from (8) is considered. One of the aim of our theoretical analysis is to show that, under this exact sampling assumption, we are able to sample efficiently from a close approximation of π in high-dimensional settings involving a large number of data. To this end, we have to ensure that each conditional sampling step involved in Algorithm 1 can be performed efficiently. This is established in Proposition 1 below which shows that, if ρ is sufficiently small, sampling $\mathbf{z}_i$ given $\boldsymbol{\theta}$ can be performed using rejection sampling with $\mathcal{O}(1)$ expected evaluations of U_i and its gradient.

Proposition 1 (Complexity of rejection sampling) *For any $i \in [b]$, suppose that U_i is M_i-gradient Lipschitz for some $M_i > 0$ and that U_i is m_i-strongly convex for some $m_i \geq 0$ (possibly zero). Let*

$$V_i(\mathbf{z}_i) := U_i(\mathbf{z}_i) + \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2}{2\rho^2},$$

$\mathbf{z}_i^*(\boldsymbol{\theta})$ be the unique minimizer of V_i, and $\tilde{\mathbf{z}}_i(\boldsymbol{\theta})$ be another point (an approximation of $\mathbf{z}_i^*(\boldsymbol{\theta})$). We let

$$\tilde{A}_i = \frac{1}{\rho^2} + m_i + \frac{\|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{2d_i} - \sqrt{\frac{\|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^4}{4d_i^2} + \frac{(1/\rho^2 + m_i)\,\|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{d_i}}\ ,$$

and set $\nu_{\boldsymbol{\theta}}(\mathbf{z}_i) := \mathcal{N}(\mathbf{z}_i; \tilde{\mathbf{z}}_i(\boldsymbol{\theta}), (\tilde{A}_i)^{-1} \cdot \mathbf{I}_{d_i})$.

Suppose that we take samples $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ from $\nu_{\boldsymbol{\theta}}$, and accept them with probability

$$\mathbb{P}(\mathbf{Z}_j \text{ is accepted}) = \exp\left(-\frac{\|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{2(1/\rho^2 + m_i - \tilde{A}_i)} - [V_i(\mathbf{Z}_j) - V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))] + \frac{\tilde{A}_i\|\mathbf{Z}_j - \tilde{\mathbf{z}}_i(\boldsymbol{\theta})\|^2}{2}\right).$$

Then, these accepted samples are distributed according to $\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$. Moreover, the expected number of samples taken until one is accepted is equal to

$$E_i := \left(\frac{1/\rho^2 + M_i}{\tilde{A}_i}\right)^{d_i/2} \cdot \exp\left[\frac{\|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|^2}{2}\left(\frac{1}{1/\rho^2 + m_i - \tilde{A}_i} - \frac{1}{1/\rho^2 + M_i}\right)\right]\ , \quad (9)$$

which is less than or equal to 2 if

$$\rho^2(2d_i(M_i - m_i) - m_i) \leq 1 \quad \text{and} \quad \|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\| \leq \frac{2}{7} \cdot \frac{\sqrt{1/\rho^2 + m_i}}{\sqrt{d_i}} \ . \tag{10}$$

Proof The proof is postponed to Appendix C.4. ■

Remark 2 *The choice of the approximate minimizer $\tilde{\mathbf{z}}_i(\boldsymbol{\theta})$ that we are using in our implementation is built via a few steps of gradient descent started from $\tilde{\mathbf{z}}_i^{[0]}(\boldsymbol{\theta}) = \mathbf{A}_i\boldsymbol{\theta}$, with step size $\frac{1}{1/\rho^2+M_i}$, that is for $j \geq 1$,*

$$\tilde{\mathbf{z}}_i^{[j]}(\boldsymbol{\theta}) = \tilde{\mathbf{z}}_i^{[j-1]}(\boldsymbol{\theta}) - \nabla V_i(\tilde{\mathbf{z}}_i^{[j-1]}(\boldsymbol{\theta})) \cdot \frac{1}{1/\rho^2 + M_i} \ .$$

We stop once the condition $\|\nabla V_i(\tilde{\mathbf{z}}_i^{[j]}(\boldsymbol{\theta}))\| \leq \frac{2}{7} \cdot \frac{\sqrt{1/\rho^2+m_i}}{\sqrt{d_i}}$ is satisfied, and set $\tilde{\mathbf{z}}_i$ to $\tilde{\mathbf{z}}_i^{[j]}$. Since the condition number of the function V_i equals $\kappa_i = \frac{1+\rho^2 M_i}{1+\rho^2 m_i}$, and the gradient descent decreases the norm of the gradient by a factor of $1 - 1/\kappa_i$ at each iteration, it follows that we need at most

$$\left\lceil \frac{\log \|\nabla V_i(\mathbf{A}_i\boldsymbol{\theta})\| - \log\left(\frac{2}{7} \cdot \frac{\sqrt{1/\rho^2+m_i}}{\sqrt{d_i}}\right)}{\log(1/(1-1/\kappa_i))} \right\rceil$$

iterations before stopping.

Proposition 1 shows that if $\rho^2 \leq 1/(2d_iM_i)$, then one can use rejection sampling to sample efficiently from (8). We would like to emphasize that this condition on ρ^2 is not limiting if our goal is to sample from a close approximation of π using Algorithm 1. Indeed, it follows from Propositions 7 and 8 that the bias between π_ρ defined in (6) and π in total variation is of the order $\mathcal{O}(\rho^2)\sum_{i=1}^b d_iM_i$. Hence, if we want to ensure that the bias in total variation is at most ϵ, for $\epsilon > 0$, ρ^2 has to be chosen of the order $\mathcal{O}(\epsilon)/(\sum_{i=1}^b d_iM_i)$ which is more restrictive than (10) in Proposition 1.

2.3.2 Sampling the Master Parameter

Regarding the master parameter $\boldsymbol{\theta}$, it follows from elementary calculations that the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{z}_{1:b}$ is Gaussian, that is

$$\Pi_\rho(\boldsymbol{\theta}|\mathbf{z}_{1:b}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_{1:b}), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \ , \tag{11}$$

with $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \rho^2(\sum_{i=1}^b \mathbf{A}_i^\top\mathbf{A}_i)^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^\top\mathbf{A}_i)^{-1}\sum_{i=1}^b \mathbf{A}_i^\top\mathbf{z}_i$. To ensure that this normal distribution is non-degenerate, the block matrix $[\mathbf{A}_1^\top \ldots \mathbf{A}_b^\top]$ must have full row rank. The matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is constant across iterations so its Cholesky decomposition, necessary to sample from (11), can be pre-computed in a preliminary step. In cases where the cost of Cholesky decomposition becomes prohibitive (for instance, in high-dimensional scenarios),

a lot of methods have been proposed to sample exactly or approximately from a given Gaussian distribution (Vono et al., 2021b). For instance, we could use samplers inspired from numerical linear algebra such as conjugate-gradient and Lanczos samplers (Ilic et al., 2004; Parker and Fox, 2012; Chow and Saad, 2014).

2.4 Connections with Optimization Methods

The SGS whose main steps are described in Algorithm 1 can be related to common optimization approaches. More precisely, it can be seen as the stochastic counterpart of alternating minimization (AM) algorithms based on the classical quadratic penalty method (Nocedal and Wright, 2006, Chapter 7). Instead of minimizing a given composite objective function, these algorithms transform this unconstrained minimization problem into a constrained one via a so-called variable splitting technique. This constraint is then relaxed by adding a "seemingly naive" quadratic term to the initial objective function before performing alternating minimization. In the sequel, we detail such an optimization approach and draw connections between the latter and Algorithm 1.

Quadratic penalty method. We consider the maximum a posteriori estimation problem under the posterior distribution π in (2), that is

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{b} U_i(\mathbf{A}_i\boldsymbol{\theta}) \ . \tag{12}$$

Similarly to direct sampling from π, solving directly this minimization problem might be computationally demanding because the objective function is a sum of b composite terms, the presence of linear operators acting on $\boldsymbol{\theta}$, non-differentiability or a possible distributed architecture. To bypass these issues, some authors (Wang et al., 2008; Afonso et al., 2010; van Leeuwen and Herrmann, 2015) proposed to build on variable splitting by introducing a set of auxiliary variables $\{\mathbf{z}_i\}_{i\in[b]}$ to reformulate (12) into the constrained minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\theta}\in\mathbb{R}^d, \mathbf{z}_1\in\mathbb{R}^{d_1},\ldots,\mathbf{z}_b\in\mathbb{R}^{d_b}} \sum_{i=1}^{b} U_i(\mathbf{z}_i) \\ &\text{subject to} \quad \mathbf{z}_i = \mathbf{A}_i\boldsymbol{\theta}, i \in [b] \ . \end{aligned}$$

The constraint $\mathbf{z}_i = \mathbf{A}_i\boldsymbol{\theta}$ is then relaxed by adding a quadratic penalty term in the objective function. This yields the approximate joint minimization problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d, \mathbf{z}_1\in\mathbb{R}^{d_1},\ldots,\mathbf{z}_b\in\mathbb{R}^{d_b}} U(\boldsymbol{\theta}, \mathbf{z}_{1:b}) \coloneqq \sum_{i=1}^{b} U_i(\mathbf{z}_i) + \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2} \ .$$

This optimization problem can be solved by alternating minimization (Beck, 2015). For fixed $\boldsymbol{\theta} \coloneqq \boldsymbol{\theta}^{[t-1]}$, one first minimizes $U(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ w.r.t. $\mathbf{z}_i$ for each factor $i \in [b]$ before minimizing, for fixed $\mathbf{z}_{1:b} \coloneqq \mathbf{z}_{1:b}^{[t]}$, $U(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ w.r.t. $\boldsymbol{\theta}$. Similarly to SGS and at the price of an approximation, the main benefit of this approach is that the minimization problems w.r.t. each auxiliary variable now only involve the sum of a single potential U_i without any operator and a quadratic term.

SGS and quadratic penalty methods. Interestingly, these AM steps stand for the deterministic counterpart of the conditional sampling steps in Algorithm 1. Indeed, instead of drawing a random variable following each conditional, these minimization steps only find the mode associated to each conditional probability distribution and can be related to iterated conditional modes in image processing (Besag, 1986). This shows another interesting link between optimization and simulation and complements earlier connections between these two fields. For instance, we can mention the celebrated one-to-one equivalence between gradient descent and discretized Langevin dynamics (Roberts and Tweedie, 1996; Pereyra, 2016; Durmus et al., 2018) and more recently the use of Hamiltonian dynamics to define first-order descent schemes achieving linear convergence (Duane et al., 1987; Maddison et al., 2018).

Connections and differences with ADMM. Similar to SGS and quadratic penalty approaches introduced above, the alternating direction method of multipliers (ADMM) also builds on a variable splitting trick to ease an inference task, see Boyd et al. (2011) for a recent comprehensive overview. However, the connection between SGS and ADMM stops here. Indeed, contrary to SGS and quadratic penalty methods, ADMM resorts to the so-called augmented Lagrangian and as such involves some dual variables $\mathbf{u}_{1:b}$ in the quadratic penalty terms and their iterative updates via dual ascent steps, see Algorithm 2.

Algorithm 2: Alternating Direction Method of Multipliers (ADMM)

Input: Potentials U_i for $i \in [b]$, penalty parameter ρ, initialization $\boldsymbol{\theta}^{[0]}$, $\mathbf{u}_{1:b}^{[0]}$ and nb. of iterations T.

for $t \leftarrow 1$ **to** T **do**

  for $i \leftarrow 1$ **to** b **do**

$$\mathbf{z}_i^{[t]} = \arg\min_{\mathbf{z}_i} U_i(\mathbf{z}_i) + \frac{1}{2\rho^2} \left\| \mathbf{z}_i - \mathbf{A}_i \boldsymbol{\theta}^{[t-1]} + \mathbf{u}_i^{[t-1]} \right\|^2$$

  end

$$\boldsymbol{\theta}^{[t]} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\rho^2} \sum_{i=1}^{b} \left\| \mathbf{z}_i^{[t]} - \mathbf{A}_i \boldsymbol{\theta} + \mathbf{u}_i^{[t-1]} \right\|^2$$

  for $i \leftarrow 1$ **to** b **do**

$$\mathbf{u}_i^{[t]} = \mathbf{u}_i^{[t-1]} + \mathbf{z}_i^{[t]} - \mathbf{A}_i \boldsymbol{\theta}^{[t]}$$

  end

end

3. Quantitative Results on the Bias of the Approximate Model

In order to establish explicit non-asymptotic mixing time bounds for SGS described in Algorithm 1, we will first provide in this section quantitative bounds on the bias between π_ρ and π in both total variation and 1-Wasserstein distances.

3.1 Results

To prove these non-asymptotic results, we shall introduce various regularity conditions listed in Assumption 1.

Assumption 1 (General assumptions)

(A_0) *For any $i \in [b]$, $U_i : \mathbb{R}^{d_i} \to \mathbb{R}$ is Borel measurable, $\inf_{\mathbf{z}_i \in \mathbb{R}^{d_i}} U_i(\mathbf{z}_i) > -\infty$, and $\exp(-U^\rho(\boldsymbol{\theta}))$ is integrable on $\mathbb{R}^d$ (U^ρ was defined in Equation 7).*

(A_1) *For any $i \in [b]$, U_i is L_i-Lipschitz, that is there exists $L_i \geq 0$ such that $|U_i(\mathbf{z}_i') - U_i(\mathbf{z}_i)| \leq L_i \left\|\mathbf{z}_i' - \mathbf{z}_i\right\|$, $\forall \mathbf{z}_i, \mathbf{z}_i' \in \mathbb{R}^{d_i}$.*

(A_2) *For any $i \in [b], \mathbf{z}_i \in \mathbb{R}^{d_i}$, U_i is twice continuously differentiable and $-M_i \mathbf{I}_d \preceq \nabla^2 U_i(\mathbf{z}_i) \preceq M_i \mathbf{I}_d$.*

(A_3) *For any $i \in [b]$, U_i is convex, that is for any $\alpha \in [0,1]$, $\mathbf{z}_i, \mathbf{z}_i' \in \mathbb{R}^{d_i}$, we have $U_i(\alpha \mathbf{z}_i + (1-\alpha)\mathbf{z}_i') \leq \alpha U_i(\mathbf{z}_i) + (1-\alpha) U_i(\mathbf{z}_i')$.*

(A_4) *For any $i \in [b]$, U_i is m_i-strongly convex, that is there exists $m_i \geq 0$ such that $U_i(\mathbf{z}_i) - \frac{m_i \|\mathbf{z}_i\|^2}{2}$ is convex.*

(A_5) *$d_1 = \ldots = d_b = d$ and $\mathbf{A}_1 = \ldots = \mathbf{A}_b = \mathbf{I}_d$.*

(A_6) *For any $i \in [b]$, U_i is centered, that is $\nabla U_i(\mathbf{A}_i \boldsymbol{\theta}^\star) = \mathbf{0}_d$, where $\boldsymbol{\theta}^\star$ is the global minimum of U.*

If some potentials U_i do not verify (A_6), one can first find the global minimum $\boldsymbol{\theta}^\star$ using optimization (typically it takes only a small number of iterations to do this up to machine precision for smooth and strongly convex potentials), and perform a linear shift and define their centered version $\tilde{U}_i$ as $\tilde{U}_i(\mathbf{A}_i \boldsymbol{\theta}) = U_i(\mathbf{A}_i \boldsymbol{\theta}) - \langle \mathbf{A}_i \boldsymbol{\theta}, \nabla U_i(\mathbf{A}_i \boldsymbol{\theta}^\star) \rangle$. (A_6) is not just a technical assumption that requires additional work in implementation, without making any difference. It is not difficult to construct an example when $b = 2$, $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{I}_d$, and U_1 and U_2 are two quadratics with different covariances whose minimizers are not at the same point, but at distance D. In such situations, one can show that the bias between π_ρ and π (in Wasserstein and total variational distance) can depend strongly on D, and cannot be bounded based only on the usual smoothness and strong convexity parameters m_i, M_i. Hence centering has a beneficial effect in reducing the bias of π_ρ in such situations. In Section 5, we will see that working with the centered potentials $\tilde{U}_i$ does not increase significantly the computational complexity of SGS when rejection sampling is used to sample the auxiliary variables $\mathbf{z}_i$ conditionally upon $\boldsymbol{\theta}$.

Our first proposition provides a simple way to verify that (A_0) holds.

Proposition 3 (Sufficient conditions for integrability) *Suppose that for any $i \in [b]$, U_i is Borel measurable, and we have potentials $V_i : \mathbb{R}^{d_i} \to \mathbb{R}$ satisfying that*

1. *$\inf_{\mathbf{z}_i \in \mathbb{R}^{d_i}} V_i(\mathbf{z}_i) > -\infty$,*
2. *V_i is L_i-Lipschitz, that is $|V_i(\mathbf{z}_i) - V_i(\mathbf{z}_i')| \leq L_i \|\mathbf{z}_i - \mathbf{z}_i'\|$ for any $\mathbf{z}_i, \mathbf{z}_i' \in \mathbb{R}^{d_i}$,*
3. *V_i lower bounds U_i, that is $V_i(\mathbf{z}_i) \leq U_i(\mathbf{z}_i)$ for any $\mathbf{z}_i \in \mathbb{R}^{d_i}$,*

4. $\exp\left(-\sum_{i\in[b]} V_i(\mathbf{A}_i\boldsymbol{\theta})\right)$ *is integrable on* $\mathbb{R}^d$.

Then (A_0) *holds.*

Remark 4 *It is easy to check that the Lipschitz conditions are satisfied by the potential terms* $\{U_i\}_{i\in[b]}$ *of the logistic regression, when there is no Gaussian prior. So in this case,* $\exp(-U^\rho(\boldsymbol{\theta}))$ *is integrable whenever* $\exp(-U(\boldsymbol{\theta}))$ *is integrable. If* $\min_{i\in[b]} \inf_{\mathbf{z}_i\in\mathbb{R}^{d_i}} U_i(\mathbf{z}_i) > -\infty$*, and there is a positive definite Gaussian term of dimension* d *(typically the prior) among the* U_i*s, then this again can be easily lower bounded by a Lipschitz function satisfying the conditions of Proposition 3, hence* (A_0) *holds. An alternative sufficient condition has been proposed in Proposition 1 of Plassier et al. (2021).*

Proof The proof is postponed to Appendix A.1. ■

The following proposition establishes the ergodicity of SGS.

Proposition 5 (Integrability and Ergodicity of SGS) *Under* (A_0), $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ *defines a joint probability density function* $\pi_\rho(\boldsymbol{\theta}) \propto \exp(-U^\rho(\boldsymbol{\theta}))$*, and SGS is* π_ρ*-irreducible and aperiodic.*

Proof The proof is postponed to Appendix A.1. ■

We now provide results giving non-asymptotic bounds on the bias between π_ρ and π. Only assuming a Lipschitz continuity property on the individual potential functions $\{U_i\}_{i\in[b]}$, Proposition 6 shows that this bias is of the order $\mathcal{O}(\rho\sum_i \sqrt{d_i})$ when ρ is sufficiently small. This result requires neither differentiability nor convexity and covers standard loss functions used in statistical machine learning such as Huber, pinball or logistic losses (Vono et al., 2021a).

Proposition 6 *Suppose that* π *satisfies* (A_0)*. Let* π *and* π_ρ *be as defined in* (2) *and* (5)*. For any* $i \in [b]$*, let* U_i *satisfy* (A_1)*. Then, for any* $\rho > 0$*, we have*

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq 1 - \prod_{i=1}^{b} \Delta_{d_i}^{(i)}(\rho) \ , \tag{13}$$

where for any $i \in [b]$*,*

$$\Delta_{d_i}^{(i)}(\rho) = \frac{D_{-d_i}(L_i\rho)}{D_{-d_i}(-L_i\rho)} \ .$$

The function D_{-d_i} *is the parabolic cylinder special function defined at the end of Section 1. In addition, for* ρ *sufficiently small, the bound* (13) *satisfies*

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} \leq 2\rho \sum_{i=1}^{b} d_i^{1/2} L_i + o(\rho) \ .$$

Proof The proof is a straightforward extension of Vono et al. (2021a, Corollary 3) and is omitted. ■

Our next result allows us to bound the TV, KL and 2-Wasserstein biases in terms of a single quantity.

Proposition 7 *Let*

$$I(U,U_\rho) := \int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}) \cdot (U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta}))_- \mathrm{d}\boldsymbol{\theta} + \left(\log\left(\frac{Z_{\pi_\rho}}{Z_\pi} \right) \right)_+ , \tag{14}$$

where $Z_\pi := \int_{\mathbb{R}^d} \exp(-U(\boldsymbol{\theta}))d\boldsymbol{\theta}$ *and* $Z_{\pi_\rho} := \int_{\mathbb{R}^d} \exp(-U^\rho(\boldsymbol{\theta}))d\boldsymbol{\theta}$ *are the normalizing constants associated with* π *and* π_ρ*, respectively. Then, we have*

$$\left\| \pi_\rho - \pi \right\|_{\mathrm{TV}} \leq I(U,U_\rho) \ , \quad \text{and} \quad D_{KL}(\pi \mid\mid \pi_\rho) \leq I(U,U_\rho) \ ,$$

that is the same bound holds for total variation distance and KL-divergence. For the 2-Wasserstein distance, assuming that U *is* m*-strongly convex for* $m > 0$*, we have*

$$W_2(\pi,\pi_\rho) \leq \sqrt{\frac{2}{m} \cdot I(U,V_\rho)} \ .$$

Proof The proof is postponed to Appendix B.1. ■

If the potentials $\{U_i\}_{i\in[b]}$ are now strongly convex and continuously differentiable with a Lipschitz-continuous gradient, the total variation bias is of order $\mathcal{O}(\rho^2 \sum_i d_i)$ for ρ sufficiently small.

Proposition 8 *Let* π *and* π_ρ *as defined in* (2) *and* (5)*, respectively. Suppose first that* π *satisfies* (A_0)*,* $b = 1$*,* $d_1 = d$*,* $\mathbf{A}_1$ *is full rank and that* (A_2) *holds (convexity is not required in this case). Let* $I(U,U_\rho)$ *be as in* (14)*. Then for any* $\rho > 0$*,*

$$I(U,U_\rho) \leq \frac{\rho^2 d M_1}{2} \ .$$

In the general multiple splitting case, suppose that Assumptions (A_0)*,* (A_2)*,* (A_4) *and* (A_6) *hold, and* $\det(\sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i) > 0$*. Then* U *is* m_U*-strongly convex for*

$$m_U = \lambda_{\min}\left(\sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i \right) .$$

Let $\mathbf{A} = [\mathbf{A}_1^\top, \ldots, \mathbf{A}_b^\top]^\top$ *(*$\mathbf{A}_1, \ldots, \mathbf{A}_b$ *are stacked one upon another), and*

$$\sigma_U^2 := \|\mathbf{A}^\top \mathbf{A}\| (\max_{i\leq b} M_i)^2 \cdot m_U^{-1} \ . \tag{15}$$

Then, for $0 < \rho^2 \leq \frac{1}{6\sigma_U^2}$*, we have*

$$I(U,U_\rho) \leq \frac{\rho^2}{2} \left(\sum_{i=1}^b d_i M_i \right) + \left(2 + \frac{3}{2} d \right) \rho^4 \sigma_U^4 \ . \tag{16}$$

Remark 9 *It is possible to reduce the bias and the constraint on ρ^2 in situations where some of the U_i's are quadratic, such as when there is a Gaussian prior (and also more generally in situations where for some indices $j \in [b]$, $e^{-U_j(\mathbf{z}_j)}$ can be written as the convolution of another function and a Gaussian density). In this case, instead of applying the algorithm on the original U, we can replace U_j by another quadratic potential U'_j such that $(U'_j)^\rho = U_j$, that is the convolution of $e^{-U'_j(\mathbf{z}_j)}$ and $\frac{1}{(2\pi\rho^2)^{d_j/2}} e^{-\|\mathbf{z}_j\|^2/(2\rho^2)}$ equals $e^{-U_j(\mathbf{z}_j)}$. Let $[b]^{n.q.}$ denote the set of indices that correspond to non-quadratic potentials. By a straightforward modification of the proof of Proposition 8 (elimination of the error terms caused by the difference between U_j and U_j^ρ for quadratics), one can show that the results hold with $\mathbf{A}$ changed to only contain $\mathbf{A}_i$ for $i \in [b]^{n.q.}$, σ_U^2 updated to $\sigma_U^2 := \|\mathbf{A}^\top \mathbf{A}\|(\max_{i\in[b]^{n.q.}} M_i)^2 \cdot m_U^{-1}$, and the final bound changed to*

$$I(U, U_\rho) \leq \frac{\rho^2}{2} \left(\sum_{i \in [b]^{n.q.}} d_i M_i \right) + \left(2 + \frac{3}{2} d\right) \rho^4 \sigma_U^4 \ .$$

This can improve the dimension dependence in situations when the number of data points is smaller than the dimension d, which is often the case for latent Gaussian models.

Proof The proof is postponed to Appendix B.1. ■

In the single splitting case corresponding to $b = 1$, Proposition 10 builds on the heat equation to derive an explicit and simple bound on the bias between π_ρ and π in 1-Wasserstein distance.

Proposition 10 *Suppose that (A_0) and (A_5) hold, $b = 1$, and U_1 is twice continuously differentiable, and satisfies*

$$U_1(\boldsymbol{\theta}) \geq a_1 + a_2 \|\boldsymbol{\theta}\|^\alpha \quad \text{and} \quad \|\nabla U_1(\boldsymbol{\theta})\| \leq a_3 + a_4 \|\boldsymbol{\theta}\|^\beta \ , \tag{17}$$

for some $a_2 > 0$, $\alpha > 0$, $\beta > 0$, $a_1, a_3, a_4 \in \mathbb{R}$. Then, we have

$$W_1(\pi, \pi_\rho) \leq \min\left(\rho\sqrt{d}, \frac{1}{2}\rho^2 \int_{\mathbb{R}^d} \|\nabla U_1(\boldsymbol{\theta})\| \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}\right) \ .$$

Moreover, if U_1 satisfies Assumptions (A_2) and (A_4) (gradient Lipschitz and strong convexity properties), then (17) holds, and we have

$$W_1(\pi, \pi_\rho) \leq \min\left(\rho\sqrt{d}, \frac{1}{2}\rho^2 \sqrt{M_1 d}\right) \ .$$

Proof The proof is given in Appendix B.2. Although we believe that a similar bound also holds for multiple splitting, unfortunately we have not found a way to obtain such a

Distance	Assumptions	Upper bound
$\|\pi_\rho - \pi\|_{\mathrm{TV}}$	(A_0), (A_1)	$\rho \sum_{i=1}^{b} 2\sqrt{d_i} L_i + o(\rho)$
	(A_0), (A_2), $b = 1$	$\frac{1}{2}\rho^2 M_1 d$
	(A_0), (A_2), (A_4), (A_6)	$\frac{1}{2}\rho^2 \sum_{i=1}^{b} M_i d_i + o(\rho^2)$
$W_1(\pi_\rho, \pi)$	(A_0), (A_2), (A_4), (A_5), $b = 1$	$\min\left(\rho\sqrt{d}, \frac{1}{2}\rho^2\sqrt{M_1 d}\right)$

Table 1: Non-asymptotic bounds given in Propositions 6, 8 and 10.

Distance	Assumptions	Upper bound
$\|\pi_{\mathrm{ULA}} - \pi\|_{\mathrm{TV}}$	(A_2), (A_3)	$\sqrt{2M_1 d} \cdot \rho$
$W_1(\pi_{\mathrm{ULA}}, \pi)$	(A_2), (A_4)	$\sqrt{2\frac{M_1}{m_1} d} \cdot \rho$

Table 2: Non-asymptotic bounds for ULA with step size $h = \rho^2$, see Theorem 12 of Durmus et al. (2019).

result (the heat equation argument used within the proof is not straightforward to adapt to the $b > 1$ case). Nevertheless, Proposition 8 provides an alternative Wasserstein bound for strongly convex U via Proposition 7 (which provides a bound not as sharp as this result in the single splitting case). ■

Table 1 summarizes the non-asymptotic bounds on the bias we have obtained. In the single splitting case with $\mathbf{A}_1 = \mathbf{I}_d$, we have the following two steps when moving from $\boldsymbol{\theta}^{[t]}$ to $\boldsymbol{\theta}^{[t+1]}$.

1. Sample $\mathbf{z}^{[t]} \sim \Pi_\rho(\mathbf{z}|\boldsymbol{\theta}^{[t]}) \propto \exp\left(-U(\mathbf{z}) - \|\mathbf{z} - \boldsymbol{\theta}^{[t]}\|^2/(2\rho^2)\right)$.

2. Sample $\boldsymbol{\theta}^{[t+1]} \sim \mathcal{N}(\mathbf{z}^{[t]}, \rho^2 \mathbf{I}_d)$.

By Taylor's expansion, we can see that in the $\rho \to 0$ limit, we have $\Pi_\rho(\mathbf{z}|\boldsymbol{\theta}^{[t]}) \approx \mathcal{N}(\boldsymbol{\theta}^{[t]} - \rho^2 \nabla U(\boldsymbol{\theta}^{[t]}), \rho^2 \mathbf{I}_d)$, and hence $\boldsymbol{\theta}^{[t+1]}$ is approximately distributed as $\mathcal{N}(\boldsymbol{\theta}^{[t]} - \rho^2 \nabla U(\boldsymbol{\theta}^{[t]}), 2\rho^2 \mathbf{I}_d)$, which corresponds to one ULA step with stepsize $h = \rho^2$. One can show that the same approximation holds in the multiple splitting case as well, after appropriate preconditioning. Thus, in the $\rho \to 0$ limit, SGS can be interpreted as another discretization of the Langevin diffusion. Hence it is natural to compare the bias bounds of Table 1 with the best available bias bounds for ULA; as far as we know, these were stated in Durmus et al. (2019). We do this in Table 2, for ULA with step size $h = \rho^2$. SGS has a significantly smaller bias than ULA both in total variation, and Wasserstein distances, suggesting a higher order

approximation. Indeed the bias is $\mathcal{O}(\rho^2)$ for 1-Wasserstein and total variation distances for SGS, while it is $\mathcal{O}(\rho)$ for both distances for ULA. This means that SGS can be seen as a discretization of the Langevin diffusion, whose stationary distribution is significantly less biased compared to ULA. The key reason why we have been able to provide these bounds is that we have access to a quite explicit form of the stationary distribution for SGS (see Equation 7 and Proposition 5), while there is no such explicit form available for ULA. Another valuable property of SGS in the single splitting case is that the posterior mean of π_ρ is the same as the posterior mean of the target π, since π_ρ is formed by the convolution of π and a Gaussian. Finally, we would like to highlight that SGS in the single splitting case can be used to sample from π without adding any bias by considering the marginal distribution of $\mathbf{z}$ under $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z})$ (which equals π) instead of $\pi_\rho(\boldsymbol{\theta})$ (Lee et al., 2021; Liang and Chen, 2021). We do not consider this alternative since our primary focus is on the multiple splitting scenario where this exact sampling approach is not possible.

3.2 Illustrations on a Toy Gaussian Model

We perform a sanity check of the tightness of the upper bounds derived in Section 3.1 on a toy Gaussian model for which a closed-form expression is available for both π_ρ and the considered statistical distances. The target distribution is chosen as a scalar Gaussian

$$\pi(\theta) = \mathcal{N}\left(\theta; \mu, \frac{\sigma^2}{b}\right),$$

where $b \geq 1$ and $\sigma > 0$. In the sequel, we set $\mu = 0$, $\sigma = 3$ and $b = 10$. To satisfy the assumptions associated to each distance (see Table 1 for a summary), we consider two splitting strategies.

Splitting strategy 1. Since the bound on $\|\pi - \pi_\rho\|_{\mathrm{TV}}$ is valid for any number of splitting operations, we set $U_i(\theta) = (2\sigma^2)^{-1}(\theta - \mu)^2$ for any $i \in [b]$. The marginal of θ under the instrumental hierarchical model in (5) has the closed-form expression

$$\pi_\rho(\theta) = \mathcal{N}\left(\theta; \mu, \frac{\sigma^2 + \rho^2}{b}\right).$$

Splitting strategy 2. As the bound in 1-Wasserstein distance has only been established for a single splitting operation, we set $U(\theta) \coloneqq U_1(\theta) = b(2\sigma^2)^{-1}(\theta - \mu)^2$. This yields

$$\pi_\rho(\theta) = \mathcal{N}\left(\theta; \mu, \frac{\sigma^2}{b} + \rho^2\right).$$

Figure 2 illustrates the bounds derived in Section 3.1 for both TV (with splitting strategy 1) and 1-Wasserstein (with splitting strategy 2) distances. The 1-Wasserstein distance has been calculated by numerical integration using the identity $W_1(\pi, \pi_\rho) = \int_{\mathbb{R}} |F(u) - F_\rho(u)| \mathrm{d}u$ where F and F_ρ are the cumulative distribution functions (c.d.f.) associated to π and π_ρ, respectively. For this simple problem, these bounds manage to achieve the correct decay in $\mathcal{O}(\rho^2)$ for small values of ρ: the quantitative bound on the 1-Wasserstein distance is particularly tight.

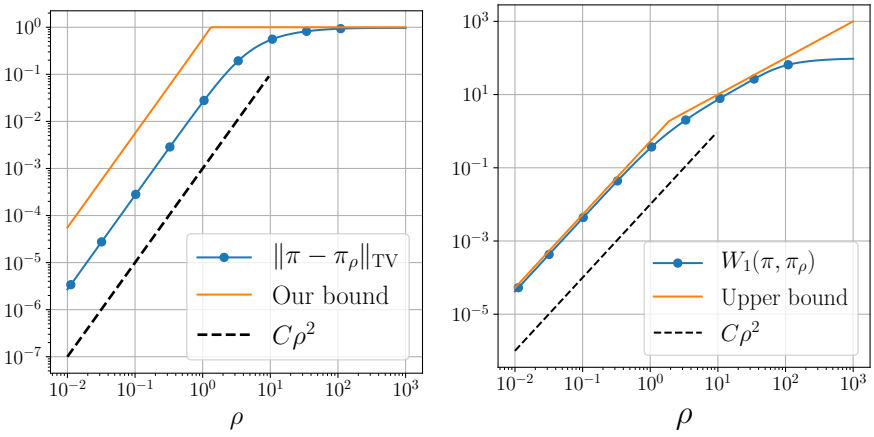

Figure 2: $\|\pi - \pi_\rho\|_{\mathrm{TV}}$ (left) and $W_1(\pi, \pi_\rho)$ (right) as a function of ρ along with the bounds established in Section 3.1 for the toy Gaussian model considered in Section 3.2. The dashed line shows the slope associated to a decay in ρ^2 in log-log scale (the parameter C stands for a positive constant).

4. Main Results: Explicit Mixing Time Bounds

We now state our main results regarding the non-asymptotic bounds on the mixing time of SGS.

4.1 Explicit Convergence Rates

In this section, we first prove a key result related to the Ricci curvature of SGS which allows us to derive explicit convergence rates for this algorithm.

4.1.1 Lower Bound on the Ricci Curvature of the SGS Kernel

The SGS sampler described in Algorithm 1 generates a Markov chain $(\boldsymbol{\theta}^{[t]})_{t\geq 1}$ of transition kernel $\mathbf{P}_{\mathrm{SGS}}$ defined by

$$\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int_{\mathbf{z}_{1:b}} \Pi_\rho(\mathbf{z}_{1:b}|\boldsymbol{\theta}) \Pi_\rho(\boldsymbol{\theta}'|\mathbf{z}_{1:b}) \mathrm{d}\mathbf{z}_{1:b} \ ,$$

where the conditional distributions associated to Π_ρ are defined in (8) and (11). For any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \mathbb{R}^d$, given a metric $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, the coarse Ricci curvature $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ of $\mathbf{P}_{\mathrm{SGS}}$, introduced by Ollivier (2009), equals

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1 - \frac{W_1^w(\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \cdot), \mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}', \cdot))}{w(\boldsymbol{\theta}, \boldsymbol{\theta}')} \ ,$$

for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \mathbb{R}^d$. We can also define this quantity for p-Wasserstein distances for any $1 \leq p \leq \infty$ by

$$K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1 - \frac{W_p^w(\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \cdot), \mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}', \cdot))}{w(\boldsymbol{\theta}, \boldsymbol{\theta}')} \ .$$

In Theorem 11, we show under Assumption (A_4) that for any $1 \leq p \leq \infty$ and a suitable metric w, $K_p(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is lower bounded by a simple quantity having an explicit dependence w.r.t. the tolerance parameter ρ and the strong convexity constants of the potential functions $\{U_i\}_{i \in [b]}$.

Theorem 11 *Suppose that π satisfies (A_0) and that (A_4) holds. Define the metric*

$$w(\boldsymbol{\theta}, \boldsymbol{\theta}') = \left\| \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i \right)^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}') \right\| \ . \tag{18}$$

Let

$$K_{\mathrm{SGS}} := 1 - \left\| \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i \right)^{-1/2} \left(\sum_{i=1}^b \frac{\mathbf{A}_i^\top \mathbf{A}_i}{1 + m_i \rho^2} \right) \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i \right)^{-1/2} \right\| \ . \tag{19}$$

Then for the transition kernel $\mathbf{P}_{\mathrm{SGS}}$ of SGS, $K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq K_{\mathrm{SGS}}$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \mathbb{R}^d$ and any $1 \leq p \leq \infty$.

Proof The proof is postponed to Appendix C.1. ■

As shown in the following corollary, Theorem 11 implies that the convergence rate of SGS towards its invariant distribution is governed by the constant K_{SGS} defined in (19).

Corollary 12 *Suppose that π satisfies (A_0) and that (A_4) holds. Then, for any $1 \leq p \leq \infty$ and any initial distribution ν on $\mathbb{R}^d$, we have*

$$\begin{aligned} W_p^w(\nu \mathbf{P}_{\mathrm{SGS}}^t, \pi_\rho) &\leq W_p^w(\nu, \pi_\rho) \cdot (1 - K_{\mathrm{SGS}})^t \ , \\ \|\nu \mathbf{P}_{\mathrm{SGS}}^t - \pi_\rho\|_{\mathrm{TV}} &\leq \mathrm{Var}_{\pi_\rho}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho} \right) \cdot (1 - K_{\mathrm{SGS}})^t \ , \end{aligned} \tag{20}$$

where W_p^w denotes the Wasserstein distance of order p w.r.t. the metric w defined in (18).

Proof The proof is postponed to Appendix C.2. ■

An attractive property of the convergence rate K_{SGS} is that it is dimension-free: it only depends on b, ρ^2 and the strong convexity parameter m_i, and neither requires differentiability nor smoothness of the potential functions $\{U_i\}_{i \in [b]}$. This is of interest since Corollary 12 can be applied to many problems where non-differentiable potential functions are considered; see Li and Lin (2010); Gu et al. (2014); Xu and Ghosh (2015).

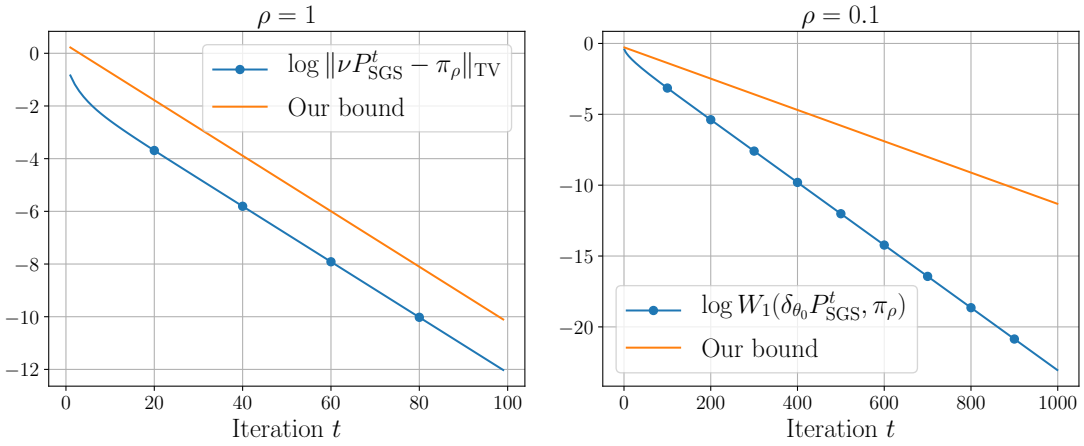

Figure 3: From left to right: $\left\|\nu P^t_{\mathrm{SGS}} - \pi_\rho\right\|_{\mathrm{TV}}$ with $\nu(\theta) = \mathcal{N}(\theta; \mu, \sigma^2/b)$ and $W_1(\delta_{\theta_0} P^t_{\mathrm{SGS}}, \pi_\rho)$ with $\theta_0 = 0$ along with the bounds shown in Theorem 11 for the toy Gaussian model considered in Section 3.2.

4.1.2 Illustrations on the Toy Gaussian Example

Before proving our mixing time bounds for the SGS, we perform a simple sanity check on the toy Gaussian example considered in Section 3.2 in order to assess the tightness of the convergence bounds stated in Corollary 12. In this case, the θ-chain follows a Gaussian auto-regressive process of order 1. We can thus compute analytically the Markov transition kernel νP^t_{SGS} and the total variation and 1-Wasserstein distances between this kernel and the invariant distribution π_ρ; see Appendix C.6. for details. For this toy Gaussian example, the convergence rate of SGS is governed by

$$K_{\mathrm{SGS}} = \frac{\rho^2}{\sigma^2 + \rho^2} \ , \text{ for the splitting strategy 1,}$$
$$K_{\mathrm{SGS}} = \frac{b\rho^2}{\sigma^2 + b\rho^2} \ , \text{ for the splitting strategy 2.}$$

Figure 3 illustrates our convergence bounds for each splitting strategy and associated statistical distance. For the total variation case, the slope in log-scale associated to our bound, which equals $\log(1 - K_{\mathrm{SGS}})$, appears to be sharp since it matches the slope associated to the observed convergence rate. Regarding the Wasserstein scenario, the slope associated to our bound is roughly equal to twice the real slope in log-scale, and hence is a bit conservative.

We are now ready to prove our main results, namely mixing time bounds associated to SGS when apply to an initial target density π which is both smooth and strongly log-concave. These assumptions will be weakened in Section 4.3.

Reference	Method	Validity	Evals
Durmus et al. (2019)	Unadjusted Langevin	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa}{\epsilon^2}\right)$
Dalalyan and Karagulyan (2019)	SGLD	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa}{\epsilon^2}\right)$
Cheng et al. (2018)	Underdamped Langevin	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa^2}{\epsilon}\right)$
Dalalyan and Riou-Durand (2020)	Underdamped Langevin	$0 < \epsilon \leq \frac{1}{\sqrt{\kappa}}$	$\mathcal{O}^*\left(\frac{\kappa^{3/2}}{\epsilon}\right)$
Chen and Vempala (2019)	Hamiltonian Dynamics	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\frac{\kappa^{3/2}}{\epsilon}\right)$
this paper	SGS with single splitting	$0 < \epsilon \leq \frac{1}{d\sqrt{\kappa}}$	$\mathcal{O}^*\left(\frac{\kappa^{1/2}}{\epsilon}\right)$

Table 3: Comparison of convergence rates in Wasserstein distance with the literature, starting from the minimizer $\boldsymbol{\theta}^\star$ of the m_1-strongly convex and M_1-smooth potential $U_1(\boldsymbol{\theta})$, with condition number $\kappa = M_1/m_1$. SGS with single splitting is implemented based on rejection sampling. $\mathcal{O}^*(\cdot)$ denotes $\mathcal{O}(\cdot)$ up to polylogarithmic factors. In the last column, the complexity stands for the number of gradient and function evaluations to get a W_1 error of $\frac{\epsilon\sqrt{d}}{\sqrt{m_1}}$. The acronym SGLD refers to stochastic gradient Langevin dynamics.

4.2 User-Friendly Mixing Time Bounds

We consider two cases, namely the single splitting strategy where $b = 1$ and the multiple one where the initial density π involves $b \geq 1$ composite potential functions. In both cases, we derived explicit expressions for the mixing time of SGS and compared them to the ones recently obtained in the MCMC literature including results associated to common subsampling MCMC approaches.

4.2.1 Single Splitting Strategy

We begin by considering the case $b = 1$ corresponding to a single splitting operation of the potential function $U \coloneqq U_1$. Since sampling from the conditional (8) is as difficult as sampling from the initial target π, this scheme is not particularly relevant from a practical point of view. Nevertheless, the convergence analysis of the scheme and its comparison with state-of-the-art MCMC approaches are still worth studying from a theoretical point of view. Indeed, the non-asymptotic results we derive in the single splitting case are simpler and allow practitioners to have a first theoretical understanding of the convergence behavior of Algorithm 1 in high-dimensional settings. By combining our error bounds on $W_1(\pi, \pi_\rho)$ from Proposition 10 to the convergence bound (20) in Corollary 12, we obtain the following complexity result.

Theorem 13 (Complexity bound for Wasserstein distance) *Suppose that π satisfies (A_0). Suppose that $b = 1$ and that Assumptions (A_2), (A_4) and (A_5) hold. Let $\boldsymbol{\theta}^\star$ be the unique minimizer of $U_1(\boldsymbol{\theta})$ and let $\nu = \delta_{\boldsymbol{\theta}^\star}$ be the initial distribution. Suppose that $\epsilon \leq 1$.*

Then, with the choice

$$\rho^2 = \max\left(\frac{\epsilon^2}{4m_1}, \frac{\epsilon}{\sqrt{m_1 M_1}}\right) , \tag{21}$$

and number of iterations $t \geq t_{\text{mix}}(\epsilon\sqrt{d/m_1}; \nu)$ *where*

$$t_{\text{mix}}(\epsilon\sqrt{d/m_1}; \nu) = \frac{\log\left(\frac{3}{\epsilon}\right)}{\log\left(1 + \max\left(\frac{\epsilon^2}{4}, \epsilon\sqrt{\frac{m_1}{M_1}}\right)\right)} , \tag{22}$$

we have

$$W_1(\nu P^t_{\text{SGS}}, \pi) \leq \frac{\epsilon}{\sqrt{m_1}}\sqrt{d} .$$

This implies that, using t *steps of SGS, we can obtain a sample that has a Wasserstein distance from the target* π *at most equal to* $\frac{\epsilon\sqrt{d}}{\sqrt{m_1}}$.

Proof The proof is postponed to Appendix C.3. ■

Several comments can be made on the result stated in Theorem 13. The expressions of both the choice of the tolerance parameter (21) and the mixing time (22) are simple and can be computed in practice. These nice properties along with the explicit dependencies of the mixing time of SGS w.r.t. the condition number $\kappa := M_1/m_1$ of U_1 and the desired precision ϵ make Theorem 13 of particular interest for practitioners. In addition, under smoothness and strong convexity of the potential U_1 (see Assumption 1), one can show that $W_1(\delta_{\boldsymbol{\theta}^\star}, \pi) \leq \sqrt{d/m_1}$ (Durmus and Moulines, 2019, Proposition 1). This quantity can be interpreted as the typical deviation associated to the sampling problem. Under the assumptions of Theorem 13, it follows that $W_1(\nu P^t_{\text{SGS}}, \pi)$ is upper bounded by ϵ times this typical deviation. Note that considering the relative precision $\epsilon\sqrt{d/m_1}$ yields a mixing time bound (22) which is invariant to the scaling of U (that is replacing U by αU with $\alpha > 0$).

For a fixed condition number κ and a sufficiently small precision ϵ, (22) implies that the mixing time of SGS scales as $\mathcal{O}(\sqrt{\kappa}\epsilon^{-1}\log(3\epsilon^{-1}))$. To be competitive with other MCMC algorithms, such as those based on Langevin or Hamiltonian dynamics, we have to ensure that the auxiliary variable $\mathbf{z}_1$ can be efficiently drawn at each iteration of Algorithm 1. In Proposition 1 in Section 2.3.1, we established that this is possible by showing that if $\epsilon \leq 1/(d\sqrt{\kappa})$, then sampling $\mathbf{z}_1$ given $\boldsymbol{\theta}$ can be performed by rejection sampling with $\mathcal{O}(1)$ expected evaluations of U_1 and its gradient. Based on this rejection sampling scheme, Table 3 compares our complexity result for SGS with single splitting with the ones derived recently in the literature. It shows that SGS compares favourably to standard MCMC methods when $0 < \epsilon \leq 1/(d\sqrt{\kappa})$ including the commonly-used subsampling approach called stochastic gradient Langevin dynamics (SGLD) (Welling and Teh, 2011). An explanation for this improved performance in terms of precision ϵ and condition number κ is the fact that SGS admits a stationary distribution with an explicit form, which we were able to exploit to establish smaller bounds on the bias for the same step size.

Theorem 14 (Complexity bound for TV distance, single splitting) *Suppose that* $b = 1$, $d_1 = d$, $\mathbf{A}_1$ *is invertible, and that Assumptions* (A_0), (A_2) *and* (A_4) *hold, with* $m_1 > 0$. *Let* $\boldsymbol{\theta}^\star$ *be the unique minimizer of* $\boldsymbol{\theta} \mapsto U(\boldsymbol{\theta}) = U_1(\mathbf{A}_1\boldsymbol{\theta})$. *Let* $\nu(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, (M_1\mathbf{A}_1^\top\mathbf{A}_1)^{-1})$ *be the initial distribution. Then for any* $0 < \epsilon \leq 1$, *with the choice*

$$\rho^2 \leq \frac{\epsilon}{dM_1} \ ,$$

and number of iterations $t \geq t_{\mathrm{mix}}(\epsilon; \nu)$ *where*

$$t_{\mathrm{mix}}(\epsilon; \nu) = \frac{\log\left(\frac{2}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}} \ ,$$

for $K_{\mathrm{SGS}} = \frac{m_1\rho^2}{1+m_1\rho^2}$ *and*

$$C = \frac{5d}{8} + \frac{d}{2}\log\left(\frac{M_1}{m_1}\right) \ ,$$

we have

$$\|\nu P_{\mathrm{SGS}}^t - \pi\|_{\mathrm{TV}} \leq \epsilon \ .$$

This means that starting from ν, *after* t *step of SGS, we are at a TV-distance at most* ϵ *from* π.

Proof The proof is postponed to Appendix C.5. ■

4.2.2 Multiple Splitting Strategy

In this section, we consider the general case where $b \geq 1$ potential functions have been split as in (5). For this scenario, the following theorem states explicit mixing time bounds in total variation distance.

Theorem 15 (Complexity bound for TV distance, multiple splitting) *Assume that* (A_0), (A_2), (A_4) *and* (A_6) *hold, and* $\det\left(\sum_{i=1}^b m_i\mathbf{A}_i^\top\mathbf{A}_i\right) > 0$. *Let* $\boldsymbol{\theta}^\star$ *be the unique minimizer of* $U(\boldsymbol{\theta}) = \sum_{i=1}^b U_i(\mathbf{A}_i\boldsymbol{\theta})$. *Let*

$$\nu(\boldsymbol{\theta}) := \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, \left(\sum_{i=1}^b M_i\mathbf{A}_i^\top\mathbf{A}_i\right)^{-1}\right)$$

be the initial distribution. Then for any $0 < \epsilon \leq 1$, *with the choice*

$$\rho^2 \leq \frac{\sum_{i=1}^b d_i M_i \left(\sqrt{1 + 8\epsilon\sigma_U^4\left(2 + \frac{3}{2}d\right)\left(\sum_{i=1}^b d_i M_i\right)^{-2}} - 1\right)}{4\sigma_U^4\left(2 + \frac{3}{2}d\right)} \wedge \frac{1}{6\sigma_U^2} \tag{23}$$

and number of iterations $t \geq t_{\mathrm{mix}}(\epsilon;\nu)$ *where*

$$t_{\mathrm{mix}}(\epsilon;\nu) = \frac{\log\left(\frac{2}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}} \ , \tag{24}$$

for K_{SGS} *defined in* (19)*, and*

$$C = d\sigma_U^2 + \rho^4 (2+d)\, \sigma_U^4 + \frac{17}{32} \sum_{i=1}^{b} d_i + \frac{1}{2} \log \left(\frac{\det\left(\sum_{i=1}^{b} M_i \mathbf{A}_i^\top \mathbf{A}_i\right)}{\det\left(\sum_{i=1}^{b} m_i \mathbf{A}_i^\top \mathbf{A}_i\right)} \right) \ ,$$

we have

$$\|\nu P_{\mathrm{SGS}}^t - \pi\|_{\mathrm{TV}} \leq \epsilon \ .$$

This means that starting from ν*, after* t *step of SGS, we are at a TV-distance at most* ϵ *from* π*.*

Proof The proof is postponed to Appendix C.5. ■

Again, note that both the tolerance parameter (23) and the ϵ-mixing time (24) are explicit and can be computed in practice. If we denote the condition number of the potential U by $\kappa := M/m$, this theorem implies that $t_{\mathrm{mix}}(\epsilon;\nu)$ scales as $\mathcal{O}(d^2\kappa/\epsilon)$ up to polylogarithmic factors. In this scenario, Table 4 compares our complexity results for SGS implemented using rejection sampling with existing results in the literature. For the same initialization ν, we have better dependencies than ULA w.r.t. both κ, d and ϵ. However, MALA seems to have better convergence rates in total variation distance in general, except in badly conditioned situations, where the rates for SGS can be better. Moreover, compared to MALA and ULA, SGS with multiple splitting is amenable to distributed and parallel computations. In this distributed environment, the complexity results shown in this table suggest that SGS is an attractive approach to sample from a smooth, strongly log-concave and composite target distribution.

4.3 Non-Strongly Log-Concave Target Density

The complexity results shown in Section 4.2 assume that each potential $\{U_i\}_{i\in[b]}$ is strongly convex. In cases where there are $b-1$ convex potential functions and the b-th one stands for an isotropic quadratic term (coming from the prior distribution for instance), this strongly-convex assumption can be met. Indeed, one can decompose the quadratic potential into $b-1$ strongly convex terms and add each of them to each individual convex potential $\{U_i\}_{i\in[b-1]}$. Nevertheless, the strongly-convex assumption is still restrictive. In this section, we extend our explicit mixing time bound in the multiple splitting scenario to densities which are smooth (see Assumption (A_2)) but such that each individual potential $\{U_i\}_{i\in[b]}$ only satisfies the standard convexity assumption (A_3) instead of satisfying the strong convexity assumption (A_4).

Similarly to Dalalyan (2017) and Dwivedi et al. (2019), we will weaken our strongly-convex assumption (A_4) by approximating each potential U_i with a strongly convex one

Reference	Method	Validity	Evals
Durmus and Moulines (2017)	ULA, $\nu = \delta_{\boldsymbol{\theta}^\star}$	$0 \leq \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d/\epsilon^2\right)$
Cheng and Bartlett (2018); Durmus et al. (2019)	ULA, $\nu = \nu_m$	$0 < \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d/\epsilon^2\right)$
Dalalyan (2017)	ULA, $\nu = \nu_M$	$0 \leq \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d^3/\epsilon^2\right)$
Durmus et al. (2019)	SGLD, $\nu = \nu_M$	$0 \leq \epsilon \leq 1$	$\mathcal{O}^*\left(\kappa^2 d^3/\epsilon^2\right)$
Dwivedi et al. (2019)	MALA, $\nu = \nu_M$	$0 < \epsilon \leq 1$	$\mathcal{O}\left(\kappa^2 d^2 \log^{1.5}\left(\frac{\kappa}{\epsilon^{1/d}}\right)\right)$
this paper	SGS, $\nu = \nu_M$	$0 < \epsilon \leq 1$	$\mathcal{O}^*(\kappa d^2/\epsilon)$

Table 4: Comparison of convergence rates in TV distance with the literature, starting from a Gaussian distribution centered at the minimizer $\boldsymbol{\theta}^\star$ of the m-strongly convex and M-smooth potential $U(\boldsymbol{\theta})$, with condition number $\kappa = \frac{M}{m}$. SGS is implemented based on rejection sampling. $\mathcal{O}^*(\cdot)$ denotes $\mathcal{O}(\cdot)$ up to polylogarithmic factors, $\nu_m(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, \frac{\mathbf{I}_d}{m})$ and $\nu_M(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, \frac{\mathbf{I}_d}{M})$. The notation ν stands for the initialization of each method.

and then applying our previous proof techniques to this approximation. More precisely, instead of the initial target density π in (2), we now consider the approximate density $\tilde{\pi}(\boldsymbol{\theta}) \propto \exp(-\tilde{U}(\boldsymbol{\theta}))$ with

$$\tilde{U}(\boldsymbol{\theta}) = \sum_{i=1}^{b} U_i(\mathbf{A}_i \boldsymbol{\theta}) + \frac{\lambda}{2} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^\star \right\|^2, \tag{25}$$

where $\lambda > 0$ and $\boldsymbol{\theta}^\star$ stands for a minimizer of U. This approximation allows us to apply Theorems 14 and 15 with the new smooth and strongly-convex constants $\tilde{M}_i = M_i + \lambda$ and $\tilde{m}_i = \lambda$ in order to find the minimum number of SGS steps such that the TV distance from $\tilde{\pi}$ is less than ϵ. To achieve an ϵ TV-distance from the initial target density π, we have to consider an additional error term to bound, namely $\|\pi - \tilde{\pi}\|_{\mathrm{TV}}$. If $\int_{\mathbb{R}^d} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^\star \right\|^4 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq d^2 R^2$ with $R > 0$, then with the choice $\lambda = 4\epsilon/(3bdR)$, we have $\|\pi - \tilde{\pi}\|_{\mathrm{TV}} \leq \epsilon/3$ (Dalalyan, 2017, Lemma 3). Combining this result with Theorem 14, the following corollary states a complexity result for the single splitting strategy. The one corresponding to the multiple splitting one can be obtained using Theorem 15 in a similar manner but is omitted here for simplicity.

Corollary 16 (Complexity bound for TV distance, no strong convexity) *Suppose that $b = 1$, Assumptions (A_0), (A_2) and (A_3) hold and*

$$\int_{\mathbb{R}^d} \left\| \boldsymbol{\theta} - \boldsymbol{\theta}^\star \right\|^4 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq d^2 R^2$$

for some $R > 0$. *Let* $\tilde{\pi}$ *be defined as in* (25). *Let* $\nu(\boldsymbol{\theta}) := \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, (\tilde{M}_1 \mathbf{A}_1^\top \mathbf{A}_1)^{-1})$ *be the initial distribution with* $\tilde{M}_1 = M_1 + \lambda$. *Then for any* $0 < \epsilon \leq 1$, *with the choices* $\lambda = 4\epsilon/(3dR)$ *and*

$$\rho^2 \leq \frac{2\epsilon}{3d(M_1 + \lambda)} \ ,$$

and number of iterations

$$t \geq \frac{\log\left(\frac{3}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}} \ ,$$

for

$$K_{\mathrm{SGS}} = \frac{\lambda\rho^2}{1 + \lambda\rho^2} \ \textit{and} \ C = \frac{5d}{8} + \frac{d}{2}\log\left(\frac{M_1 + \lambda}{\lambda}\right) \ ,$$

we have

$$\|\nu P_{\mathrm{SGS}}^t - \pi\|_{\mathrm{TV}} \leq \epsilon \ .$$

This means that starting from ν, *after* t *step of SGS applied to the approximate density* $\tilde{\pi}$, *we are at a TV-distance at most* ϵ *from* π.

Proof The proof is straightforward. It follows from the triangle inequality and Theorem 14. ■

Compared to our mixing time bound derived under the assumption that the potential U_1 is strongly convex, Corollary 16 shows that relaxing the strongly convex assumption affects negatively the dependence w.r.t. both the dimension d and the precision ϵ, as it scales as $\mathcal{O}^*(M_1 d^2/\epsilon^2)$. Nevertheless, this complexity result improves upon that in Dalalyan (2017); Dwivedi et al. (2019) for the unadjusted Langevin algorithm (ULA) and the Metropolized random walk (MRW), which respectively scale as $\mathcal{O}^*(M_1^2 d^3/\epsilon^4)$ and $\mathcal{O}^*(M_1^2 d^3/\epsilon^2)$.

4.4 Comparison with Existing Divide-and-Conquer and Subsampling-Based MCMC Schemes

So far, we have mainly compared the theoretical behavior of SGS in high-dimensional scenarios with common MCMC schemes such as those derived from Langevin and Hamiltonian dynamics, see Tables 3 and 4. In this section, we discuss and compare when possible the theoretical results associated to SGS with those associated to existing divide-and-conquer and subsampling-based MCMC approaches. Regarding divide-and-conquer approaches, although a lot of algorithms have been proposed over the past ten years (see Section 1), very few non-asymptotic and explicit convergence results exist up to the authors' knowledge (Plassier et al., 2021). Among available results, we can cite those associated to non-parametric approaches proposed by Wang and Dunson (2013); Neiswanger et al. (2014); Wang et al. (2015) which showed that these methods scaled exponentially with respect to the dimension d because of the use of kernel density estimates. For general subsampling-based approaches which do not resort to the Bernstein-von Mises approximation, explicit

bounds have been recently derived for (variance-reduced) stochastic gradient MCMC algorithms such as SGLD (Dalalyan and Karagulyan, 2019; Durmus et al., 2019). As illustrated in Tables 3 and 4, our non-asymptotic theoretical analysis shows that SGS is competitive with other state-of-the-art MCMC algorithms.

5. Numerical Illustrations

This section aims at illustrating the main theoretical results of Section 4. We consider three different examples which satisfy all the assumptions required in our main statements. The first experiment considers the case where the target π is a multivariate Gaussian density while the second one sets π to be a mixture of two multivariate Gaussian densities. Finally, the third experiment considers a Bayesian binary logistic regression problem with a Gaussian prior. For all approaches and experiments, the initial distribution will be set to $\nu = \mathcal{N}(\boldsymbol{\theta}^\star, (\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i)^{-1})$ for the TV distance and to $\nu = \delta_{\boldsymbol{\theta}^\star}$ for the 1-Wasserstein one. Although SGS is amenable to a distributed implementation (Rendell et al., 2021), all the experiments have been run on a serial computer to emphasize that it is even beneficial in this context. The experiments have been carried out on a Dell Latitude 7390 laptop equipped with an Intel(R) Core(TM) i5-8250U 1.60 GHz processor, with 16.0 GB of RAM, running Windows 10.

5.1 Multivariate Gaussian Density

In this example, we want to verify empirically the dependencies of the mixing times derived in Section 4 w.r.t. the dimension d, the desired precision ϵ and the condition number κ of the potential U. We consider a target zero-mean Gaussian density on $\mathbb{R}^d$

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{Q}\boldsymbol{\theta}\right),$$

where $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is a positive definite precision matrix. In the sequel, $\mathbf{Q}$ will be chosen to be diagonal and anisotropic, that is $\mathbf{Q} = \mathrm{diag}(q_1, \ldots, q_d)$, with $q_i \neq q_j$ for $i \neq j$. The resulting potential function $U := U_1 = \boldsymbol{\theta}^\top \mathbf{Q}\boldsymbol{\theta}/2$ is strongly convex and smooth with parameters $m = \min_{i\in[d]} q_i$ and $M = \max_{i\in[d]} q_i$. Since computing the total variation distance between continuous and multidimensional measures is challenging, we discretized the latter over a set of bins and consider the error between the empirical marginal densities associated to the least favorable direction, that is along the eigenvector associated to m. In the following, we will illustrate our mixing time results for both 1-Wasserstein and total variation distances in the strongly log-concave case.

Dimension dependence. We set here $\epsilon = 0.1$, $m = 1/4$, $M = 1$ such that $\kappa := M/m = 4$ and are interested in the dimension dependence of our ϵ-mixing time result for SGS. We let the dimension d vary between 10^1 and 10^3 and ran SGS for each case. We measured its ϵ-mixing time by recording the smallest iteration such that the discrete total variation error falls below the desired precision ϵ. The mixing time has been averaged over 10 independent runs. Figure 4 illustrates the behavior of the mixing time of SGS w.r.t. the dimension d in log-log scale. In order to assess the dimension dependency, we performed a linear fit and reported the slope of the linear model. According to Table 4, the dimension dependence is

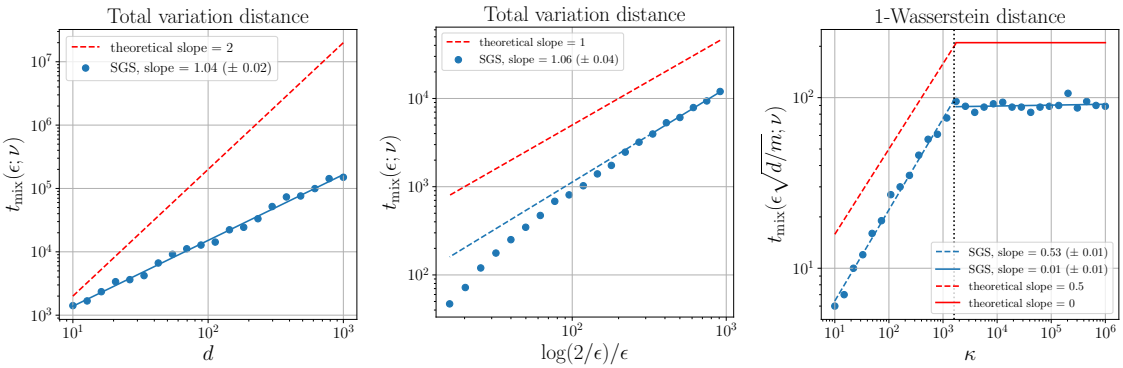

Figure 4: Multivariate Gaussian. (left and middle) ϵ-mixing times for the total variation distance and (right) $\epsilon\sqrt{d/m}$-mixing times for the 1-Wasserstein distance.

of order $\mathcal{O}(d^2)$. Interestingly, we found in this example that the dimension dependence of the mixing time of SGS is linear w.r.t. d.

Precision dependence. We set here $d = 2$ and $\kappa = 3$ while the prescribed precision ϵ varies between 6×10^{-3} and 1.6×10^{-1}, and ran SGS for each case. As before, we measured its ϵ-mixing time by recording the smallest iteration such that the discrete total variation error falls below the desired precision ϵ. Figure 4 shows the behavior of the mixing time of SGS w.r.t. $\log(2/\epsilon)\epsilon^{-1}$ in log-log scale. For sufficiently small precisions, this figure confirms our theoretical result which states that the mixing time of SGS scales as $\mathcal{O}(\log(2/\epsilon)\epsilon^{-1})$.

Condition number dependence. Regarding the 1-Wasserstein distance and the complexity results depicted in Table 3, the main difference between existing MCMC approaches is the dependence w.r.t. the condition number κ of the potential function U. Here, we aim at verifying the latter numerically. To this purpose, we set $d = 10$, $\epsilon = 0.1$ and let κ vary between 10^1 and 10^6. From (22), it appears that the dependence of the mixing time of SGS depends on $\max\{\epsilon^2/4, \epsilon/\sqrt{\kappa}\}$. This quantity equals $\epsilon/\sqrt{\kappa}$ for $\kappa \leq 1600$ and $\epsilon^2/4$ otherwise. Hence, we are expecting to retrieve a dependence in $\kappa^{1/2}$ for small and moderate κ and a mixing time only depending on ϵ for larger values of the condition number. We performed 50 independent runs of SGS and stopped them when their empirical Wasserstein error fell below $\epsilon\sqrt{d/m}$. The results are depicted on Figure 4 in log-log scale. As before, we did a linear fit to assess the dependency of the mixing time w.r.t. the condition number κ. The slope of the linear model for SGS equals 0.53 for $\kappa \leq 1600$ (depicted with a black dotted vertical line) which confirms the theoretical dependence of the order $\mathcal{O}(\kappa^{1/2})$. As expected, the mixing time of SGS becomes independent of κ for larger values.

5.2 Gaussian Mixture

In this second experiment, also considered by Dalalyan (2017) and Dwivedi et al. (2019), we show that the values of the tolerance parameter ρ and the mixing time $t_{\mathrm{mix}}(\epsilon;\nu)$ recommended by Theorem 14 indeed yield approximate samples having a distribution close to π. We also verify that the running time required to generate such samples is reasonable, and

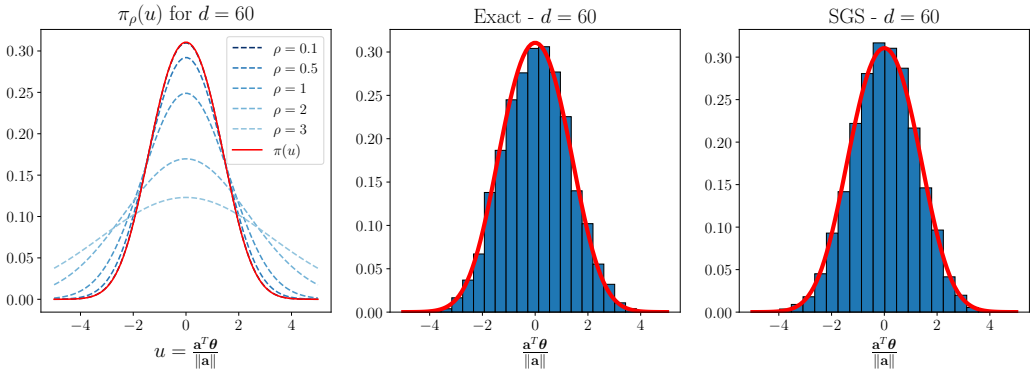

Figure 5: Gaussian mixture with $d = 60$. From left to right: behavior of $\pi_\rho(u)$ w.r.t. ρ with $u = \mathbf{a}^\top\boldsymbol{\theta}/\|\mathbf{a}\|$; empirical distribution obtained by exact sampling from π; empirical distribution obtained by sampling from π_ρ with the guidelines recommended in Theorem 15. The histograms have been computed using 2500 independent samples and the precision has been set to $\epsilon = 0.1$. In all figures, the red curve stands for $\pi(u)$.

compare it to the running time of ULA to achieve the same prescribed precision ϵ. To this purpose, let us consider the simple problem of generating samples from a mixture of two Gaussian densities with density π defined, for all $\boldsymbol{\theta} \in \mathbb{R}$, by

$$\pi(\boldsymbol{\theta}) = \frac{1}{2(2\pi)^{d/2}} \left(\exp\left(-\frac{\|\boldsymbol{\theta} - \mathbf{a}\|^2}{2} \right) + \exp\left(-\frac{\|\boldsymbol{\theta} + \mathbf{a}\|^2}{2} \right) \right)$$
$$\propto \exp\left(-U(\boldsymbol{\theta})\right),$$

where

$$U(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{a}\|^2 - \log\left(1 + e^{-2\boldsymbol{\theta}^\top \mathbf{a}}\right),$$

and $\mathbf{a} \in \mathbb{R}^d$ is a fixed vector involved in the mean of each Gaussian density. If $\|\mathbf{a}\| < 1$, one can show that U is M-smooth and m-strongly convex with $m = 1 - \|\mathbf{a}\|^2$ and $M = 1$. In the sequel, we choose $\mathbf{a}$ such that $\|\mathbf{a}\| = 1/\sqrt{2}$, which also implies that the global minimizer of U is $\boldsymbol{\theta}^\star = \mathbf{0}_d$. Since π admits a finite second order moment, all the assumptions required in Theorem 14 are verified. We now consider a single splitting strategy on U leading to the joint approximate density $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z})$ defined in (5) with $b = 1$ and $\mathbf{A}_1 = \mathbf{I}_d$. Under this distribution, the marginal density $\pi_\rho(\boldsymbol{\theta})$ writes

$$\pi_\rho(\boldsymbol{\theta}) = \frac{1}{2(2\pi(1+\rho^2))^{d/2}} \left(\exp\left(-\frac{\|\boldsymbol{\theta} - \mathbf{a}\|^2}{2(1+\rho^2)} \right) + \exp\left(-\frac{\|\boldsymbol{\theta} + \mathbf{a}\|^2}{2(1+\rho^2)} \right) \right),$$

Dimension d	4	8	12	16	20	30	40	60
$t_{\text{mix}}(\epsilon;\nu)$ ($\times 10^3$) for SGS	3	10	23	40	62	138	244	548
$t_{\text{mix}}(\epsilon;\nu)$ ($\times 10^3$) for ULA	29	87	184	330	532	1,350	2,729	7,742
Efficiency of SGS w.r.t. ULA	10.8	8.6	8.2	8.3	8.6	9.8	11.1	14.1
CPU time [s] for SGS	1	7	29	62	114	335	749	2,416
CPU time [s] for ULA	6	31	135	302	589	1,974	4,766	15,096
Efficiency of SGS w.r.t. ULA	5.6	4.6	4.7	4.9	5.2	5.9	6.4	6.2

Table 5: Gaussian mixture. Comparison between SGS and ULA for a prescribed precision $\epsilon = 0.1$. For SGS, $t_{\text{mix}}(\epsilon;\nu)$ has been computed by using Theorem 14 while for ULA, the mixing time bound derived in Dalalyan (2017, Corollary 1) has been used. CPU time information corresponds to the running time necessary to draw 10^3 independent samples having a distribution at most ϵ total variation distance from π.

and simply corresponds to a mixture of the two initial Gaussian densities but with respective variance now inflated by a factor ρ^2. The one-dimensional approximate density $\pi_\rho(u)$ of $u = \mathbf{a}^\top \boldsymbol{\theta} / \|\mathbf{a}\|$ is depicted in Figure 5 for $d = 60$ and compared to the true target $\pi(u)$.

Illustrations of Theorem 14. We now illustrate the guidelines for ρ and the number of iterations t, stated in Theorem 14, to achieve an ϵ-error in total variation distance. To this purpose, we set $\epsilon = 0.1$, $d = 60$ and launched 2500 independent runs of SGS. The conditional distribution of $\mathbf{z}$ given $\boldsymbol{\theta}$ is a mixture of two Gaussians with common covariance matrix $\boldsymbol{\Sigma} = \rho^2/(1+\rho^2)\mathbf{I}_d$, respective mean vectors $\boldsymbol{\mu}_1 = (\boldsymbol{\theta} + \mathbf{a}\rho^2)/(1+\rho^2)$ and $\boldsymbol{\mu}_2 = (\boldsymbol{\theta} - \mathbf{a}\rho^2)/(1+\rho^2)$ and respective weights $w_1 = 1$ and $w_2 = \exp(-4\boldsymbol{\theta}^\top \mathbf{a}/(2(1+\rho^2))$. We can sample exactly from this mixture by first drawing a Bernoulli random variable B with probability $p = w_1/(w_1+w_2)$ and then setting $\mathbf{z} = B(\boldsymbol{\xi} + \boldsymbol{\mu}_1) + (1-B)(\boldsymbol{\xi} + \boldsymbol{\mu}_2)$ where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{\Sigma})$. In order to assess the relevance of the samples generated with SGS, we generated 2500 independent samples directly from π by an exact sampler similar to the one used to sample $\mathbf{z}$. To provide an illustration of the quality of the samples drawn with SGS, we computed the one-dimensional projection $u = \mathbf{a}^\top \boldsymbol{\theta} / \|\mathbf{a}\|$ and showed its empirical distribution in Figure 5. The empirical distribution of the samples drawn using SGS is indeed close to π and is visually indistinguishable from the one of the exact samples.

Computational complexity of SGS. We now verify empirically the computational complexity of SGS, that is the number of iterations and the overall running time for generating samples with some prescribed precision ϵ. We compare this complexity to that of ULA (Dalalyan, 2017). Starting from the same initial distribution $\nu = \mathcal{N}(\boldsymbol{\theta}^\star, M^{-1}\mathbf{I}_d)$ and with $\epsilon = 0.1$, Table 5 reports the number of iterations $t_{\text{mix}}(\epsilon;\nu)$ required, in theory, to obtain a sample whose distribution is at most ϵ in total variation from π and the CPU time needed to generate 10^3 such samples. For ULA, $t_{\text{mix}}(\epsilon;\nu)$ has been computed by using the mixing time bound derived in Dalalyan (2017, Corollary 1). We observe that for $d \in [4, 60]$, both

the number of iterations and the running time for generating 10^3 independent samples with SGS are much smaller than that of ULA.

This second experiment confirms our theoretical statement that SGS is able to generate accurate samples for a reasonable computational budget compared to popular alternatives such as ULA.

5.3 Bayesian Binary Logistic Regression

The previous two sections illustrated our theoretical results for a single splitting strategy and $\mathbf{A}_1 = \mathbf{I}_d$. In this section, we consider a more challenging problem, namely Bayesian binary logistic regression. This model involves $b > 1$ potential functions, matrices $\{\mathbf{A}_i\}_{i\in[b]}$ not equal to the identity, and is such that the observations might be distributed over a set of b nodes within a cluster. As introduced in Section 2, SGS is of interest for this scenario since it allows to sample from the posterior distribution of interest in such distributed environments. In the sequel, we will also show the benefits of splitting $\mathbf{A}_i\boldsymbol{\theta}$ instead of only splitting the variable of interest $\boldsymbol{\theta}$ as in Vono et al. (2019); Rendell et al. (2021). This goal will be conducted by illustrating numerically our mixing time bounds and assessing the efficiency of the rejection sampling procedure (see Proposition 1) used to sample the auxiliary variables $\mathbf{z}_{1:b}$, in both cases.

5.3.1 Problem Formulation

As introduced in Example 2, the logistic regression problem considers a set of observed data $\{\mathbf{x}_i, y_i\}_{i\in[n]}$ where the binary labels $y_i \in \{0, 1\}$ are related to the unknown regression parameter $\boldsymbol{\theta}$ via the model (3). In a Bayesian framework, a standard approach consists of assigning a zero-mean Gaussian prior to $\boldsymbol{\theta}$ with diagonal precision matrix $\boldsymbol{\Sigma}^{-1} = \tau\mathbf{I}_d$ as in (4); see Albert and Chib (1993); Holmes and Held (2006). Instead, we set here $\boldsymbol{\Sigma}^{-1} = \alpha\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, with $\alpha = 3d/(\pi^2 n)$ which corresponds to a Zellner prior (Sabanes Bove and Held, 2011; Hanson et al., 2014). Such a choice leads to a posterior density $\pi(\boldsymbol{\theta}) \propto \exp(-U(\boldsymbol{\theta}))$ with

$$U(\boldsymbol{\theta}) = \sum_{i=1}^n y_i\mathbf{x}_i^\top\boldsymbol{\theta} + \log\left[1 + \exp\left(-\mathbf{x}_i^\top\boldsymbol{\theta}\right)\right] + \frac{\alpha}{2}\left\|\mathbf{x}_i^\top\boldsymbol{\theta}\right\|^2 .$$

Sampling from this posterior density can be conducted by exploiting the mixture representation of the binomial likelihood which involves the Polya-Gamma distribution, and then performing Gibbs sampling (Polson et al., 2013). Nevertheless, although this algorithm has been shown to be uniformly ergodic w.r.t. the TV distance, the best known explicit result for its ergodicity constant degrades exponentially quickly with n and d, see Choi and Hobert (2013). We propose here to sample approximately from the posterior using SGS. We will consider and compare two splitting strategies.

Splitting strategy 1. The first strategy sets $b = n$, $\mathbf{A}_i = \mathbf{x}_i^\top$ for $i \in [b]$ and leads to the approximate posterior density (5) with

$$U_i(z_i) = y_i z_i + \log\left[1 + \exp\left(-z_i\right)\right] + \frac{\alpha}{2}z_i^2 , \quad \forall i \in [b] .$$

In this case, U_i is m_i-strongly convex and M_i-smooth with $m_i = \alpha$ and $M_i = \alpha + 1/4$, respectively, and hence verifies (A_2) and (A_4). As detailed after Proposition 8, we can verify (A_6) by centering U_i with a simple linear shift. In the sequel, we will assume that such a shifting has been performed which implies that all the assumptions in Theorem 15 are verified. The interest of this first splitting strategy is that the conditional posterior probability densities of z_i given $\boldsymbol{\theta}$ are univariate and easy to sample.

Splitting strategy 2. The second strategy mimics the one used by Rendell et al. (2021) and considers that the data $\{\mathbf{x}_i, y_i\}_{i\in[n]}$ is divided into b shards $\{\mathrm{D}_i\}_{i\in[b]}$. For simplicity, we will assume that n is a multiple of b such that $\mathrm{card}(\mathrm{D}_i) = n/b$ for all $i \in [b]$. In contrast to the first splitting strategy, we use here $\mathbf{A}_i = \mathbf{I}_d$ for $i \in [b]$. This yields

$$U_i(\mathbf{z}_i) = \sum_{j\in\mathrm{D}_i} y_j \mathbf{x}_j^\top \mathbf{z}_i + \log\left[1 + \exp\left(-\mathbf{x}_j^\top \mathbf{z}_i\right)\right] + \frac{\alpha}{2}\left\|\mathbf{x}_j^\top \mathbf{z}_i\right\|^2 \ , \quad \forall i \in [b] \ .$$

Here U_i is m_i-strongly convex and M_i-smooth with $m_i = \alpha\lambda_{\min}(\sum_{j\in\mathrm{D}_i} \mathbf{x}_j\mathbf{x}_j^\top)$ and $M_i = (\alpha+1/4)\lambda_{\max}(\sum_{j\in\mathrm{D}_i} \mathbf{x}_j\mathbf{x}_j^\top)$, where $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ stand for the smallest and largest eigenvalues of a matrix $\mathbf{M}$, respectively. As before, we assume that an appropriate centering of U_i has been performed to satisfy (A_6). In some scenarios, such a splitting strategy is expected to be less efficient than the first one for two main reasons. First, the conditional density of $\mathbf{z}_i$ given $\boldsymbol{\theta}$ is d-dimensional and sampling from it is as difficult as sampling from π. Second, the condition number $\kappa = \sum_i M_i / \sum_i m_i$ associated with this strategy (denoted κ_2) might be very large compared to the one associated to the splitting strategy 1 (denoted κ_1). Indeed, the ratio of these two condition numbers is

$$\frac{\kappa_2}{\kappa_1} = \frac{\sum_{i=1}^{b} \lambda_{\max}\left(\sum_{j\in\mathrm{D}_i} \mathbf{x}_j\mathbf{x}_j^\top\right)}{\sum_{i=1}^{b} \lambda_{\min}\left(\sum_{j\in\mathrm{D}_i} \mathbf{x}_j\mathbf{x}_j^\top\right)}. \tag{26}$$

This ratio is expected to be large when d is large and the correlation between the covariates within each group is high. The splitting strategy 1 can be thought of as a preconditioning technique whose efficiency is measured by the ratio (26).

5.3.2 Efficient Sampling with the SGS

For this experiment, also considered in Dalalyan (2017), we generated a synthetic data set $\{\mathbf{x}_i, y_i\}_{i=1}^n$ by drawing the covariates $\mathbf{x}_i$ from a Rademacher distribution before normalizing the latter such that $\|\mathbf{x}_i\| = 1$. Each binary label y_i was then drawn from a Bernoulli distribution with probability of success equal to $\sigma(\mathbf{x}_i^\top \boldsymbol{\theta}_{\mathrm{true}})$, where $\sigma(\cdot)$ is the sigmoid function and $\boldsymbol{\theta}_{\mathrm{true}} = \mathbf{1}_d$.

Mixing times. In this first sub-experiment, we compare our mixing time bounds in Theorem 15 for the two splitting strategies detailed previously. We set $\epsilon = 0.01$, $n = 1,000$ and let b vary from $b = 5$ to $b = 10$ for the splitting strategy 2. The theoretical ϵ-mixing times for the TV distance associated to the two splitting strategies are reported in Figure 6. To give an idea of the order of magnitude of these mixing times, the ones associated to

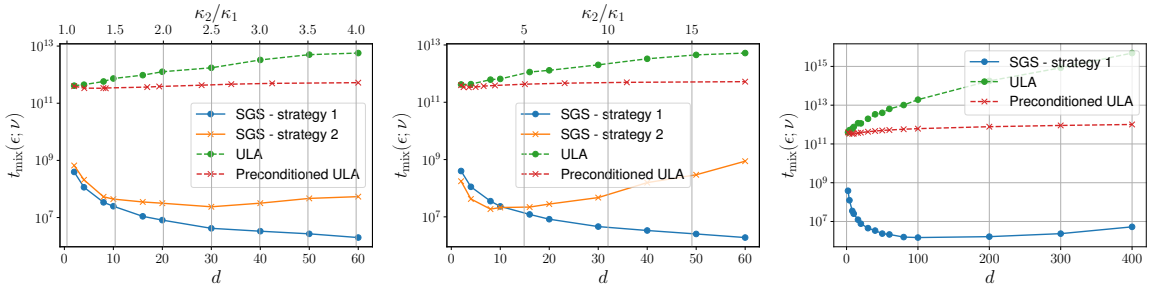

Figure 6: Logistic regression. (left and middle) Behavior of the mixing times of the two splitting strategies w.r.t. the ratio κ_1/κ_2. For strategy 2, $b = 5$ (left) and $b = 10$ (middle), while for strategy 1, $b = n = 1000$ on both figures. (right) Behavior of the splitting strategy 1 in higher dimensions.

ULA and its preconditioned version (Dalalyan, 2017) using the same starting distribution ν are also displayed. As expected, the splitting strategy 1 needs, in theory, less iterations than the splitting strategy 2 to achieve a prescribed precision ϵ when the ratio κ_2/κ_1 is large. Finally, it is clear that the mixing times of SGS are again competitive compared to the ones derived in the recent literature for other MCMC algorithms.

Efficiency of the rejection sampling scheme. In this second sub-experiment, we complement the previous analysis by showing that drawing the auxiliary variables $\mathbf{z}_i$ given $\boldsymbol{\theta}$ can indeed be conducted efficiently with rejection sampling. For each instance of SGS (with either splitting strategy 1 or 2), we used the rejection sampling scheme detailed in Proposition 1 in Section 2.3.1 to sample the auxiliary variables. To this purpose, we considered different scenarios where $d \in \{2, 10, 50\}$ and $n \in \{200, 10^3, 10^4\}$. We ran SGS over $T = 100$ iterations and averaged the number of rejection steps over these iterations for each auxiliary variable $\mathbf{z}_i$. In Table 6, we reported the largest average number of rejection steps per iteration obtained among the b auxiliary variables for each splitting strategy. For the nine different scenarios, the average number of rejection steps per iteration is near 1 which confirms the theoretical results of Proposition 1. Overall, SGS appears to be a promising and efficient approach to sample from smooth and strongly log-concave distributions.

6. Conclusion

In this paper, we have provided a detailed theoretical study of a recent and promising MCMC algorithm, namely SGS, which is amenable to a distributed implementation and shares strong similarities with quadratic penalty approaches in optimization. Under a strong log-concavity assumption, we have obtained explicit dimension-free convergence rates for this sampler under both Wasserstein and total variation distances. Combined with quantitative bounds on the bias induced by this algorithm, we have derived explicit bounds on its mixing time under reasonable assumptions which can be easily verified in practice. In addition to be amenable to distributed and parallel computations, these results showed that

d	2			10			50		
n	200	1,000	10,000	200	1,000	10,000	200	1,000	10,000
SGS 1 ($b = n$)	1.04	1.04	1.03	1.03	1.05	1.03	1.06	1.05	1.03
SGS 2 ($b = 2$)	1.13	1.16	1.34	1.26	1.08	1.16	1.06	1.27	1.03
SGS 2 ($b = 5$)	1.22	1.14	1.26	1.08	1.14	1.08	1.00	1.10	1.34
SGS 2 ($b = 10$)	1.12	1.07	1.05	1.31	1.17	1.16	1.41	1.07	1.13

Table 6: Logistic regression. Average number of samples proposed until one is accepted per iteration for SGS 1 (associated to splitting strategy 1) and SGS 2 (associated to splitting strategy 2).

SGS can compete and even improve upon standard MCMC schemes in terms of computational complexity. Our theoretical results have been supported with numerical illustrations which confirmed the efficiency of SGS even on a serial computer.

There are a few additional interesting questions to address. All our theoretical results assume that the auxiliary variables $\mathbf{z}_i$ are drawn from the exact conditional probability density at each iteration of SGS. Although this is possible for interesting models such as logistic regression, one might have to sample approximately these variables using Metropolis-Hastings or proximal MCMC scheme (Pereyra, 2016; Durmus et al., 2018; Vargas et al., 2020) in more complex scenarios and it would be interesting to extend our results to such settings. Another interesting extension would be to consider whether using a sequence $\{\rho_t\}_{t\in\mathbb{N}}$ instead of a fixed parameter ρ could be beneficial by determining convergence rates in this scenario.

Acknowledgments

This material is based upon work supported in part by the U.S. Army Research Laboratory and the U.S. Army Research Office, by the U.K. Ministry of Defence (MoD), and by the U.K. Engineering and Physical Research Council (EPSRC) under grant number EP/R013616/1. It is also supported by EPSRC grants EP/R034710/1 and EP/R018561/1. The authors thank the GdR ISIS and Rémi Bardenet from Université de Lille for funding MV's visit to Oxford through the internationalization grant "Effet tunnel". Part of this work has been supported by the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANITI). We thank Solomon Jacobs and Andreas Eberle for pointing out an error in the proof of Proposition 8 in a previous version of this paper.

Appendix Appendix A. Additional Details and Proofs for Section 2

This section aims at proving the results claimed in Section 2.

Appendix A.1. Integrability of π_ρ and Ergodicity of SGS

Proof [Proof of Proposition 5] Let $U(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ be defined as in (5), then by repeated application of Tonelli's theorem (Tonelli, 1909) to integrate out $\mathbf{z}_1, \dots \mathbf{z}_b$, we have

$$\int_{\boldsymbol{\theta},\mathbf{z}_{1:b}} \exp(-U(\boldsymbol{\theta}, \mathbf{z}_{1:b}))\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{z}_{1:b} = \int_{\boldsymbol{\theta},\mathbf{z}_{1:b}} \exp\left(-\sum_{i=1}^{b} U_i(\mathbf{z}_i) + \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{z}_{1:b}$$
$$= \int_{\boldsymbol{\theta}} \exp\left(-\sum_{i=1}^{b} U_i^\rho(\mathbf{A}_i\boldsymbol{\theta})\right)\mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \exp(-U^\rho(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}.$$

By Assumption (A_0), $\exp(-U^\rho(\boldsymbol{\theta}))$ is integrable, hence $\exp(-U(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ is also integrable, and $\Pi_\rho(\boldsymbol{\theta}, \mathbf{z}_{1:b})$ is a probability density. The π-irreducibility and aperiodicity of SGS follows because SGS defined on the extended state space including $\mathbf{z}_{1:b}$ is a Gibbs sampler with systematic scan, and it satisfies the positivity condition of Gibbs sampling (since the densities are always positive); see for instance Roberts and Smith (1994). ■

Proof [Proof of Proposition 3] Note that we have

$$\begin{aligned}\exp(-U_i^\rho(\mathbf{w}_i)) &= \int_{\mathbf{z}_i\in\mathbb{R}^d} \exp\left(-U_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{w}_i\|^2}{2\rho^2}\right)\cdot\frac{\mathrm{d}\mathbf{z}_i}{(2\pi\rho^2)^{d_i/2}}\\ &\le \int_{\mathbf{z}_i\in\mathbb{R}^d} \exp\left(-V_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{w}_i\|^2}{2\rho^2}\right)\cdot\frac{\mathrm{d}\mathbf{z}_i}{(2\pi\rho^2)^{d_i/2}}\\ &\le \exp\left(-V_i(\mathbf{w}_i)\right)\cdot\int_{\mathbf{z}_i\in\mathbb{R}^d} \exp\left(L_i\|\mathbf{z}_i - \mathbf{w}_i\| - \frac{\|\mathbf{z}_i - \mathbf{w}_i\|^2}{2\rho^2}\right)\cdot\frac{\mathrm{d}\mathbf{z}_i}{(2\pi\rho^2)^{d_i/2}}.\end{aligned}$$

It is easy to show that $L_i\|\mathbf{z}_i - \mathbf{w}_i\| - \frac{\|\mathbf{z}_i-\mathbf{w}_i\|^2}{2\rho^2} \le -\frac{\|\mathbf{z}_i-\mathbf{w}_i\|^2}{4\rho^2}$ whenever $\|\mathbf{z}_i - \mathbf{w}_i\| \ge 2\rho^2 L_i$, hence this integral is finite, and $\exp(-U_i^\rho(\mathbf{w}_i)) \le \exp\left(-V_i(\mathbf{w}_i)\right) C_i$ for some $C_i < \infty$. Hence, we have that $\exp(-U^\rho(\boldsymbol{\theta})) \le \exp\left(-\sum_{j\in[b]} V_j(\mathbf{A}_j\boldsymbol{\theta})\right)\cdot\prod_{j\in[b]} C_j$. The integrability of $\exp(-U^\rho(\boldsymbol{\theta}))$ now follows from our assumption that $\exp\left(-\sum_{j\in[b]} V_j(\mathbf{A}_j\boldsymbol{\theta})\right)$ is integrable. ■

Appendix Appendix B. Proofs for the Results of Section 3

This section gives the proofs and technical details associated to the results presented in Section 3.

Appendix B.1. Non-Asymptotic Bound for $I(U, U^\rho)$

In this section, we are going to bound the bias of the stationary distribution of SGS (π_ρ) from π. We start by the proof of Propostion 7, which shows that we can bound the total variation, KL and Wasserstein-2 distances between π_ρ and π in terms of $I(U, U_\rho)$.

Proof [Proof of Proposition 7] By using the notations $f(\boldsymbol{\theta})_- = -\min(f(\boldsymbol{\theta}), 0)$ and $f(\boldsymbol{\theta})_+ = \max(f(\boldsymbol{\theta}), 0)$, note that

$$\begin{aligned}\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} &= \frac{1}{2}\int_{\boldsymbol{\theta}\in\mathbb{R}^d} |\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})| \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_+ \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(1 - \frac{\pi_\rho(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)_+ \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{27}$$

since

$$\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_+ \mathrm{d}\boldsymbol{\theta} - \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \left(\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} = 0,$$
$$|\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta})| = (\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta}))_+ + (\pi(\boldsymbol{\theta}) - \pi_\rho(\boldsymbol{\theta}))_-.$$

Using the definitions of π and π_ρ, we have

$$\left\|\pi_\rho - \pi\right\|_{\mathrm{TV}} = \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(1 - \exp\left(U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta})\right) \cdot \frac{Z_\pi}{Z_{\pi_\rho}}\right)_+ \mathrm{d}\boldsymbol{\theta}$$

using the fact that $(1 - \exp(x))_+ \leq x_-$ for any $x \in \mathbb{R}$,

$$\leq \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\log\left(\frac{Z_\pi}{Z_{\pi_\rho}}\right) + U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} \tag{28}$$

$$\leq \left(\log\left(\frac{Z_{\pi_\rho}}{Z_\pi}\right)\right)_+ + \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(U^\rho(\boldsymbol{\theta}) - U(\boldsymbol{\theta})\right)_+ \mathrm{d}\boldsymbol{\theta} = I(U, U_\rho), \tag{29}$$

hence the TV bound follows. For KL-divergence, note that (28) satisfies that

$$D_{KL}(\pi \mid\mid \pi_\rho) = \int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}) \log\left(\frac{\pi(\boldsymbol{\theta})}{\pi_\rho(\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{\theta} \leq \int_{\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\log\left(\frac{Z_\pi}{Z_{\pi_\rho}}\right) + U^\rho(\boldsymbol{\theta}) - U(\boldsymbol{\theta})\right)_+ \mathrm{d}\boldsymbol{\theta},$$

hence this is also bounded by $I(U, U_\rho)$. Finally, the Wasserstein bound follows from the KL bound and Lemma 9 of Cheng and Bartlett (2018). ■

Now we will state a few definitions and prove some auxiliary lemmas, and then prove Proposition 8. We have $U(\boldsymbol{\theta}) = \sum_{i=1}^b U_i(\mathbf{A}_i\boldsymbol{\theta})$, and

$$\pi(\boldsymbol{\theta}) = \frac{\exp(-U(\boldsymbol{\theta}))}{Z_\pi}, \text{ for a normalising constant } Z_\pi = \int_{\boldsymbol{\theta}} \exp(-U(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}.$$

Similarly, by Proposition 5, we have

$$\pi_\rho(\boldsymbol{\theta}) = \frac{\exp(-U^\rho(\boldsymbol{\theta}))}{Z_{\pi_\rho}}.$$

The following lemma states some bounds on $U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta})$.

Lemma 17 *Let*

$$\overline{B}(\boldsymbol{\theta}) \coloneqq \frac{\rho^2}{2}\sum_{i=1}^{b}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2,$$
$$\underline{B}(\boldsymbol{\theta}) \coloneqq \sum_{i=1}^{b}\left(\frac{\rho^2}{2(1+\rho^2 M_i)}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2 - \frac{d_i}{2}\log(1+\rho^2 M_i)\right).$$

Then assuming (A_0) *and* (A_2)*, we have* $\underline{B}(\boldsymbol{\theta}) \le U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta})$*. Assuming* (A_0)*,* (A_2) *and* (A_4)*, we have* $U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta}) \le \overline{B}(\boldsymbol{\theta})$*.*

Proof First, note that

$$\exp(U(\boldsymbol{\theta}) - U^\rho(\boldsymbol{\theta})) = \exp\left(\sum_{i=1}^{b}\left(U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i^\rho(\mathbf{A}_i\boldsymbol{\theta})\right)\right).$$

From (7), it is clear that

$$\exp(U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i^\rho(\mathbf{A}_i\boldsymbol{\theta})) = \int_{\mathbf{z}_i\in\mathbb{R}^d}\exp\left(U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i(\mathbf{z}_i) - \frac{\|\mathbf{z}_i - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)\cdot\frac{\mathrm{d}\mathbf{z}_i}{(2\pi\rho^2)^{d_i/2}}. \tag{30}$$

Using (A_2), and second order Taylor expansion, for each $i\in[b]$, we have

$$U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i(\mathbf{z}_i) \ge \nabla U_i(\mathbf{A}_i\boldsymbol{\theta})^\top(\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i) - \frac{M_i}{2}\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2.$$

Hence, using (30), we have

$$\begin{aligned}
&\exp\left(\sum_{i=1}^{b}\left(U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i^\rho(\mathbf{A}_i\boldsymbol{\theta})\right)\right)\\
&\ge \prod_{i=1}^{b}(2\pi\rho^2)^{-d_i/2}\int_{\mathbf{z}_i\in\mathbb{R}^{d_i}}\exp\left(\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})^\top(\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i) - \left(\frac{1+\rho^2 M_i}{2\rho^2}\right)\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2\right)\mathrm{d}\mathbf{z}_i\\
&= \prod_{i=1}^{b}\left(\exp\left(\frac{\rho^2}{2(1+\rho^2 M_i)}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2\right)\left(\frac{1}{1+\rho^2 M_i}\right)^{d_i/2}\right) \qquad (31)\\
&= \exp\left(\sum_{i=1}^{b}\left(\frac{\rho^2}{2(1+\rho^2 M_i)}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2 - \frac{d_i}{2}\log(1+\rho^2 M_i)\right)\right) = \exp(\underline{B}(\boldsymbol{\theta})), \qquad (32)
\end{aligned}$$

hence the lower bound follows.

For the upper bound, we now use (A_4) (convexity of the individual potential functions U_i for $i\in[b]$), which by Taylor expansion yields that for every $i\in[b]$,

$$U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i(\mathbf{z}_i) \le \nabla U_i(\mathbf{A}_i\boldsymbol{\theta})^\top(\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i) - \frac{m_i}{2}\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{z}_i\|^2.$$

Then, it follows that

$$\exp(U(\boldsymbol{\theta}) - U^{\rho}(\boldsymbol{\theta})) = \exp\left(\sum_{i=1}^{b}\left(U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i^{\rho}(\mathbf{A}_i\boldsymbol{\theta})\right)\right)$$
$$\leq \prod_{i=1}^{b}(2\pi\rho^2)^{-d_i/2}\int_{\mathbf{z}_i\in\mathbb{R}^{d_i}}\exp\left(\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})^{\top}(\mathbf{A}_i\boldsymbol{\theta}-\mathbf{z}_i) - \frac{1+\rho^2 m_i}{2\rho^2}\left\|\mathbf{A}_i\boldsymbol{\theta}-\mathbf{z}_i\right\|^2\right)\mathrm{d}\mathbf{z}_i$$
$$= \prod_{i=1}^{b}\frac{1}{(1+\rho^2 m_i)^{d_i/2}}\cdot\exp\left(\sum_{i=1}^{b}\frac{\rho^2\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2}{2(1+\rho^2 m_i)}\right) \tag{33}$$
$$\leq \exp\left(\frac{\rho^2}{2}\sum_{i=1}^{b}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2\right) = \exp(\overline{B}(\boldsymbol{\theta})), \tag{34}$$

hence the upper bound follows. ■

Lemma 18 *Suppose that Assumptions* (A_2) *and* (A_4) *hold. Let*

$$\beta(\boldsymbol{\theta}) := \left(\sum_{i=1}^{b}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2\right)^{1/2},$$

and $\mathbf{A}$ *be the* $(d_1 + \ldots d_b)\times d$ *matrix created by stacking* $\mathbf{A}_1,\ldots,\mathbf{A}_b$ *one upon another, starting with* $\mathbf{A}_1$ *on the top and ending with* $\mathbf{A}_b$*. Then* β *is a Lipschitz function with respect to the Euclidean distance, with Lipschitz constant*

$$L_\beta = \|\mathbf{A}^{\top}\mathbf{A}\|^{1/2}\max_{i\leq b} M_i.$$

Proof Assuming that $\boldsymbol{\theta}\neq\boldsymbol{\theta}^{\star}$, we have $\beta(\boldsymbol{\theta})>0$, and thus $\nabla\beta(\boldsymbol{\theta})$ exists, and it has the form

$$\nabla\beta(\boldsymbol{\theta}) = \frac{\sum_{i=1}^{b}\mathbf{A}_i^{\top}\nabla^2 U_i(\mathbf{A}_i\boldsymbol{\theta})\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})}{\left(\sum_{i=1}^{b}\left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2\right)^{1/2}}.$$

Let $\mathbf{w} := (\nabla U_1(\mathbf{A}_1\boldsymbol{\theta}),\ldots,\nabla U_b(\mathbf{A}_b\boldsymbol{\theta}))\in\mathbb{R}^{d_1+\ldots+d_b}$, and $\mathbf{D} := \mathrm{diag}(\nabla^2 U_1(\mathbf{A}_1\boldsymbol{\theta}),\ldots,\nabla^2 U_b(\mathbf{A}_b\boldsymbol{\theta}))$ (a block matrix with diagonal blocks corresponding $\nabla^2 U_1(\mathbf{A}_1\boldsymbol{\theta}),\ldots,\nabla^2 U_b(\mathbf{A}_b\boldsymbol{\theta})$). Then we have

$$\nabla\beta(\boldsymbol{\theta}) = \frac{\mathbf{A}^{\top}\mathbf{D}\mathbf{w}}{\|\mathbf{w}\|},\ \text{hence}$$
$$\|\nabla\beta(\boldsymbol{\theta})\|^2 = \frac{\mathbf{w}^{\top}\mathbf{D}\mathbf{A}\mathbf{A}^{\top}\mathbf{D}\mathbf{w}}{\|\mathbf{w}\|^2}$$
$$\leq \|\mathbf{D}\|^2\|\mathbf{A}\mathbf{A}^{\top}\|$$
$$\leq \|\mathbf{A}^{\top}\mathbf{A}\|(\max_{i\leq b} M_i)^2,$$

so $\|\nabla\beta(\boldsymbol{\theta})\| \leq L_\beta$. Since this bound holds everywhere except possibly at $\boldsymbol{\theta}^\star$, the Lipschitz property is easy to show by a limiting argument. ■

The next result is a technical lemma that will be used in the proof.

Lemma 19 *Suppose that $\mu(\boldsymbol{\theta}) \propto \exp(-U(\boldsymbol{\theta}))$ is a density on $\mathbb{R}^d$ satisfying that U is twice continuously differentiable, and $\nabla^2 U(\boldsymbol{\theta}) \succeq m\mathbf{I}_d$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$ for some $m > 0$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-Lipschitz function with $\mathbb{E}_\mu(f) = 0$. Then for $0 \leq \lambda \leq \frac{m}{6L^2}$, we have*

$$\mathbb{E}_\mu\left[\mathrm{e}^{\lambda(f^2 - \mathbb{E}_\mu(f^2))^2}\right] \leq \mathrm{e}^{\frac{4\lambda^2 L^4}{m^2}}. \tag{35}$$

Proof First, we are going to assume that f is continuously differentiable. Then by the Lipschitz property, we have $\|\nabla f(\boldsymbol{\theta})\| \leq L$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$, and using Corollary 3.4 of Huang and Tropp (2021), we can bound the moments of $f^2(\boldsymbol{\theta}) - \mathbb{E}_\mu(f^2)$ as follows. For integers $p \geq 2$, by using this Corollary 3.4 twice (first on the function $f^2 - \mathbb{E}_\mu(f^2)$, and then on f in the third line), we have

$$\begin{aligned}
\mathbb{E}_\mu\left[(f^2 - \mathbb{E}_\mu(f^2))^p\right] &\leq \mathbb{E}_\mu\left[\left|f^2 - \mathbb{E}_\mu(f^2)\right|^p\right] \\
&\leq \frac{1}{m^{p/2}}(p-1)^{p/2}\mathbb{E}_\mu\left[(2\|\nabla f\|^2 f^2)^{p/2}\right] \\
&\leq \frac{1}{m^{p/2}} 2^{p/2}(p-1)^{p/2} L^p \mathbb{E}_\mu\left[|f|^p\right] \\
&\leq \frac{1}{m^p} 2^{p/2}(p-1)^p L^{2p}.
\end{aligned}$$

Therefore, using the power series representation of the exponential function, we have

$$\begin{aligned}
\mathbb{E}_\mu\left[\mathrm{e}^{\lambda(f^2 - \mathbb{E}_\mu(f^2))^2}\right] &\leq 1 + \sum_{p=2}^{\infty}\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^p \frac{(p-1)^p}{p!} \\
&\leq 1 + \frac{1}{2}\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^2 + \sum_{p=3}^{\infty}\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^p \frac{(p-1)^p}{(p/\mathrm{e})^p\sqrt{2\pi p}} \\
&\leq 1 + \frac{1}{2}\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^2 + \sum_{p=3}^{\infty}\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^p \frac{\mathrm{e}^{p-1}}{\sqrt{2\pi\cdot 3}},
\end{aligned}$$

where we have used the facts that $p! \geq (p/\mathrm{e})^p\sqrt{2\pi p}$ by Robbins (1955), and $((p-1)/p)^p \leq 1/\mathrm{e}$ for every $p \geq 1$. By summing up the geometric series, we have that for $0 \leq \frac{\mathrm{e}\sqrt{2}\lambda L^2}{m} < 1$,

$$\mathbb{E}_\mu\left[\mathrm{e}^{\lambda(f^2 - \mathbb{E}_\mu(f^2))^2}\right] \leq 1 + \frac{1}{2}\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^2 + \frac{\left(\frac{\sqrt{2}\lambda L^2}{m}\right)^3 \mathrm{e}^2}{\sqrt{6\pi}} \cdot \frac{1}{1 - \frac{\mathrm{e}\sqrt{2}\lambda L^2}{m}}.$$

It is easy to show that for $0 \leq \frac{\lambda L^2}{m} \leq \frac{1}{6}$, the above sum is bounded by $\mathrm{e}^{\frac{4\lambda^2 L^4}{m^2}}$, implying (35). Finally, the proof without assuming differentiability of f follows by using Theorem 6

of Azagra et al. (2007), and a limiting argument using the dominated convergence theorem. ■

The next result presents some bounds on certain moment generating functions.

Lemma 20 *Suppose that Assumptions* (A_0), (A_2), (A_4), *and* (A_6) *hold. Let*

$$\sigma_U^2 := L_\beta^2 \cdot m_U^{-1} = \|\mathbf{A}^\top \mathbf{A}\| (\max_{i \leq b} M_i)^2 \cdot m_U^{-1}.$$

Then for $0 \leq s \leq \frac{1}{12\sigma_U^2}$, *we have*

$$\int_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}) e^{s \cdot \beta^2(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \leq \mathrm{e}^{s\mathbb{E}_\pi(\beta^2)} \cdot \mathrm{e}^{s^2(8\sigma_U^4 + 4(\mathbb{E}_\pi(\beta))^2\sigma_U^2)}.$$

Moreover, we have $(\mathbb{E}_\pi(\beta))^2 \leq \mathbb{E}_\pi(\beta^2) \leq d\sigma_U^2$.

Proof From Assumption (A_4), we have

$$\nabla^2 U(\boldsymbol{\theta}) = \sum_{i=1}^b \mathbf{A}_i^\top \nabla^2 U_i(\mathbf{A}_i\boldsymbol{\theta}) \mathbf{A}_i \succeq \sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i \succeq \lambda_{\min}\left(\sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i\right) \mathbf{I}_d = m_U \mathbf{I}_d. \quad (36)$$

By Theorem 5.2 of Ledoux (2001), π satisfies a log-Sobolev inequality with constant $C := m_U^{-1}$. Therefore, by Herbst's argument (see equation (5.8) on page 95 of Ledoux (2001)), and the L_β-Lipschitz property of the function β shown in Lemma 18, for every $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_\pi(\mathrm{e}^{\lambda(\beta - \mathbb{E}_\pi(\beta))}) \leq \mathrm{e}^{C\lambda^2 L_\beta^2/2}. \quad (37)$$

By the decomposition $\beta^2(\boldsymbol{\theta}) = \mathbb{E}_\pi(\beta)^2 + (\beta(\boldsymbol{\theta}) - \mathbb{E}_\pi(\beta))^2 + 2\mathbb{E}_\pi(\beta)(\beta(\boldsymbol{\theta}) - \mathbb{E}_\pi(\beta))$, we have

$$\mathbb{E}_\pi(\mathrm{e}^{s\beta^2}) = \mathrm{e}^{s\mathbb{E}_\pi(\beta)^2} \cdot \mathbb{E}_\pi\left(\mathrm{e}^{s(\beta - \mathbb{E}_\pi(\beta))^2} \cdot \mathrm{e}^{2s\mathbb{E}_\pi(\beta)(\beta - \mathbb{E}_\pi(\beta))}\right)$$

by the Cauchy-Schwarz inequality

$$\leq \mathrm{e}^{s\mathbb{E}_\pi(\beta)^2} \cdot \left[\mathbb{E}_\pi\left(\mathrm{e}^{2s(\beta - \mathbb{E}_\pi(\beta))^2}\right)\right]^{1/2} \cdot \left[\mathbb{E}_\pi\left(\mathrm{e}^{4s\mathbb{E}_\pi(\beta)(\beta - \mathbb{E}_\pi(\beta))}\right)\right]^{1/2}.$$

Lemma 19 implies that for $0 \leq s \leq \frac{1}{12\sigma_U^2}$, we have

$$\mathbb{E}_\pi\left(\mathrm{e}^{2s(\beta(\boldsymbol{\theta}) - \mathbb{E}_\pi(\beta))^2}\right) \leq \mathrm{e}^{2s\mathbb{E}_\pi[(\beta(\boldsymbol{\theta}) - \mathbb{E}_\pi(\beta))^2]} \cdot \mathrm{e}^{16s^2\sigma_U^4}. \quad (38)$$

By (37) for $\lambda = 4s\mathbb{E}_\pi(\beta)$, we have

$$\mathbb{E}_\pi\left(\mathrm{e}^{4s\mathbb{E}_\pi(\beta)(\beta - \mathbb{E}_\pi(\beta))}\right) \leq \mathrm{e}^{8s^2 C(\mathbb{E}_\pi(\beta))^2 L_\beta^2}.$$

The first claim of the lemma now follows by rearrangement. Additionally we have $(\mathbb{E}_\pi(\beta))^2 \leq \mathbb{E}_\pi(\beta^2)$, which can be further bounded as

$$\mathbb{E}_\pi(\beta^2) = \mathbb{E}_\pi(\sum_{i=1}^{b} \|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\|^2)$$

using Assumptions (A_2) and (A_6)

$$\begin{aligned}
&\leq \mathbb{E}_\pi(\sum_{i=1}^{b} M_i^2 \|\mathbf{A}_i(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\|^2) \\
&\leq (\max_{1\leq i\leq b} M_i)^2 \|\mathbf{A}^\top\mathbf{A}\| \mathbb{E}_\pi(\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|^2) \\
&\leq (\max_{1\leq i\leq b} M_i)^2 \|\mathbf{A}^\top\mathbf{A}\| d m_U^{-1} = d\sigma_U^2,
\end{aligned}$$

where we have used the fact that $\mathbb{E}_\pi(\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|^2) \leq \frac{d}{m_U}$ by Proposition 1 part (ii) of Durmus and Moulines (2019). ■

Lemma 21 *If we assume that $b = 1$, $d_1 = d$, and $\mathbf{A}_1$ is of full rank, then we have $Z_\pi = Z_{\pi_\rho}$ for any ρ. More generally, assume that (A_0), (A_2), $(A_4)$and (A_6) hold. Then for $\rho^2 \leq \frac{1}{6\sigma_U^2}$,*

$$\log\left(\frac{Z_\pi}{Z_{\pi_\rho}}\right) \geq -\mathbb{E}_\pi(\overline{B}(\boldsymbol{\theta})) - \rho^4(2+d)\sigma_U^4. \tag{39}$$

Proof Firstly, we that $b = 1$, $d_1 = d$, and $\mathbf{A}_1$ is of full rank. Then one can show that

$$\begin{aligned}
&\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp(-U_1^\rho(\mathbf{A}_1\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}\in\mathbb{R}^d}\int_{\mathbf{z}_1\in\mathbb{R}^d} \exp\left(-U_1(\mathbf{z}_1) - \frac{\|\mathbf{z}_1-\mathbf{A}_1\boldsymbol{\theta}\|^2}{2\rho^2}\right)\cdot\frac{\mathrm{d}\mathbf{z}_1}{(2\pi\rho^2)^{d/2}}\mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp(-U_1(\mathbf{A}_1\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}.
\end{aligned}$$

Hence, in this case $Z_\pi = Z_{\pi_\rho}$.

Now we look at the general multiple splitting case. Note that using the fact that $\int_{\boldsymbol{\theta}\in\mathbb{R}^d}\pi(\boldsymbol{\theta})\left(1-\frac{\pi_\rho(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}\right)\mathrm{d}\boldsymbol{\theta} = 0$, it follows that

$$\frac{Z_\pi}{Z_{\pi_\rho}} = \left(\int_{\boldsymbol{\theta}\in\mathbb{R}^d}\pi(\boldsymbol{\theta})\exp\left(\sum_{i=1}^{b}\left(U_i(\mathbf{A}_i\boldsymbol{\theta}) - U_i^\rho(\mathbf{A}_i\boldsymbol{\theta})\right)\right)\mathrm{d}\boldsymbol{\theta}\right)^{-1}. \tag{40}$$

By (34) and (40), we have

$$\frac{Z_\pi}{Z_{\pi_\rho}} \geq \left(\int_{\boldsymbol{\theta}\in\mathbb{R}^d}\pi(\boldsymbol{\theta})\exp(\overline{B}(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta}\right)^{-1}.$$

Note that $\overline{B}(\boldsymbol{\theta}) = \frac{\rho^2}{2}\beta(\boldsymbol{\theta})^2$, hence by Lemma 20, we have that for $\rho^2 \leq \frac{1}{6\sigma_U^2}$,

$$\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \exp(\overline{B}(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta} \leq \mathrm{e}^{\mathbb{E}_\pi(\overline{B}(\boldsymbol{\theta}))} \cdot \mathrm{e}^{2\rho^4\sigma_U^4 + \rho^4\mathbb{E}_\pi(\boldsymbol{\theta})^2\sigma_U^2} \leq \mathrm{e}^{\mathbb{E}_\pi(\overline{B}(\boldsymbol{\theta}))} \cdot \mathrm{e}^{(2+d)\rho^4\sigma_U^4}.$$

The claim now follows by rearrangement. ■

Now we are ready to prove our bias bound.

Proof [Proof of Proposition 8] In the single splitting case, assuming (A_0), $b = 1$, $d_1 = d$, and that $\mathbf{A}_1$ is invertible, by Lemma 21, we have $\frac{Z_\pi}{Z_{\pi_\rho}} = 1$. Combining this and (32) with our bound (28), we obtain that

$$\begin{aligned}
I(U, U^\rho) &\leq \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\underline{B}(\boldsymbol{\theta})\right)_- \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(\frac{\rho^2 \left\|\nabla U_1(\mathbf{A}_1\boldsymbol{\theta})\right\|^2}{2(1+\rho^2 M_1)} - \frac{d}{2}\log(1+\rho^2 M_1) \right)_- \mathrm{d}\boldsymbol{\theta} \\
&\leq \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi(\boldsymbol{\theta}) \left(-\frac{d}{2}\log(1+\rho^2 M_1) \right)_- \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{d}{2} M_1 \rho^2.
\end{aligned}$$

In the general case, note that the m_U-strong convexity follows by (36). We have the lower bound (39) on $\log\left(\frac{Z_\pi}{Z_{\pi_\rho}}\right)$. By combining this and (32) with our bound (28), we obtain that for $\rho^2 \leq \frac{1}{6\sigma_U^2}$, we have

$$\begin{aligned}
&I(U, U^\rho) \\
&\leq \sum_{i=1}^b \frac{d_i}{2}\log(1+\rho^2 M_i) + \sum_{i=1}^b \frac{\rho^4 M_i}{2(1+\rho^2 M_i)} \mathbb{E}_\pi(\|\nabla U_i(A_i\boldsymbol{\theta})\|^2) + (2+d)\rho^4\sigma_U^4 \\
&\leq \frac{\rho^2}{2}(\sum_{i=1}^b d_i M_i) + \frac{\rho^4}{2}(\max_{1\leq i\leq b} M_i)\mathbb{E}_\pi(\beta^2) + (2+d)\rho^4\sigma_U^4 \\
&\leq \frac{\rho^2}{2}(\sum_{i=1}^b d_i M_i) + \left(2 + \frac{3}{2}d\right)\rho^4\sigma_U^4,
\end{aligned}$$

where in the last step we have used the facts that $\mathbb{E}_\pi(\beta^2) \leq \sigma_U^2 d$ (by Lemma 20) and that $\max_{1\leq i\leq b} M_i \leq \sigma_U^2$ (by the definition of σ_U^2). ■

Appendix B.2. Non-Asymptotic Bounds for the 1-Wasserstein Distance

The result shown in Proposition 10 shows a Wasserstein error rate bound in the single splitting case ($b = 1$). Its proof is given below.

Proof [Proof of Proposition 10] The integrability of U_1 follows from assumption (17). Assume without loss of generality that $U_1(\boldsymbol{\theta})$ is normalised, i.e. $\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp(-U_1(\boldsymbol{\theta}))\mathrm{d}\boldsymbol{\theta} = 1$ (if it is not, we can add to the potential an appropriate constant). Then the distribution

$$\pi_\rho(\boldsymbol{\theta}) = \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(-U_1(\mathbf{z}) - \frac{\|\boldsymbol{\theta}-\mathbf{z}\|^2}{2\rho^2}\right)\mathrm{d}\mathbf{z}$$

is the convolution of $\pi(\boldsymbol{\theta}) = \exp(-U_1(\boldsymbol{\theta}))$ and a d-dimensional Gaussian random variable with mean zero and covariance $\rho^2\mathbf{I}_d$. In particular, it is clear that

$$\begin{aligned}\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} &= \int_{\mathbf{z}\in\mathbb{R}^d} \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\boldsymbol{\theta}\in\mathbb{R}^d} \exp\left(-U_1(\mathbf{z}) - \frac{\|\boldsymbol{\theta}-\mathbf{z}\|^2}{2\rho^2}\right)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{z} \\ &= \int_{\mathbf{z}\in\mathbb{R}^d} \exp(-U_1(\mathbf{z})) = 1.\end{aligned}$$

The first part of the bound follows from the fact that the expectation of the norm of this Gaussian random variable is bounded by $\rho\sqrt{d}$ (since the expectation of the square of the norm is $\rho^2 d$, this follows by Jensen's inequality).

In order to obtain the second part, we are going to use the dual formulation of the 1-Wasserstein distance (see e.g. Remark 6.5 of Villani (2008)),

$$\begin{aligned}W_1(\pi,\pi_\rho) &= \sup_{g:\|g\|_{\mathrm{Lip}}\leq 1} \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} - \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \sup_{g\in C^1(\mathbb{R}^d):\|\nabla g\|_\infty\leq 1} \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} - \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \qquad (41)\end{aligned}$$

where the second equality follows from the fact that differentiable functions g with $\|\nabla g\|_\infty \leq 1$ are dense among 1-Lipschitz functions on $\mathbb{R}^d$.

The evolution of a density π_ρ as we increase the variance ρ^2 is known to follow the heat equation, see Section 2.4 of Lawler (2010),

$$\frac{\mathrm{d}}{\mathrm{d}(\rho^2)}\pi_\rho(\boldsymbol{\theta}) = \frac{1}{2}\Delta\pi_\rho(\boldsymbol{\theta}),$$

where $\Delta\pi_\rho(\boldsymbol{\theta}) = \sum_{i=1}^d \frac{\partial^2}{\partial\theta_i^2}\pi_\rho(\boldsymbol{\theta})$ denotes the Laplacian of π_ρ. By integration, we obtain that

$$\sup_{g\in C^1(\mathbb{R}^d):\|\nabla g\|_\infty<1} \frac{\mathrm{d}}{\mathrm{d}(\rho^2)} \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \sup_{g\in C^1(\mathbb{R}^d):\|\nabla g\|_\infty\leq 1} \frac{1}{2}\int_{\boldsymbol{\theta}\in\mathbb{R}^d} g(\boldsymbol{\theta})\Delta\pi_\rho(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Now if we define the functional

$$\mathcal{F}(\mu) := \sup_{g\in C^1(\mathbb{R}^d):\|\nabla g\|_\infty\leq 1} \frac{1}{2}\int_{\boldsymbol{\theta}\in\mathbb{R}^d} g(\boldsymbol{\theta})\Delta\mu(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

Then it is easy to see that this is convex ($\mathcal{F}(\alpha\mu + (1-\alpha)\nu) \leq \alpha\mathcal{F}(\mu) + (1-\alpha)\mathcal{F}(\nu)$ for $\alpha\in[0,1]$) and shift-invariant (if $\nu(\mathbf{x}) = \mu(\mathbf{x}-\mathbf{a})$ some constant $\mathbf{a}\in\mathbb{R}^d$, then $\mathcal{F}(\nu) = \mathcal{F}(\mu)$). Therefore it follows by the argument on pages 1-2 of Bennett and Bez (2015) (monotonicity property of the heat semigroup for convex functionals) that $\mathcal{F}(\pi_\rho) \leq \mathcal{F}(\pi)$ for every $\rho \geq 0$.

Initially, we have

$$\mathcal{F}(\pi) = \sup_{g \in C^1(\mathbb{R}^d): \|\nabla g\|_\infty \leq 1} \frac{1}{2} \sum_{i=1}^{d} \int_{\boldsymbol{\theta} \in \mathbb{R}^d} g(\boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i^2} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

After separating $\boldsymbol{\theta}$ to $\theta_i \in \mathbb{R}$ and $\boldsymbol{\theta}_{-i} \in \mathbb{R}^{d-1}$ (denoting the rest of the coordinates), we have

$$\begin{aligned}
&\int_{\boldsymbol{\theta} \in \mathbb{R}^d} g(\boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i^2} \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}_{-i} \in \mathbb{R}^{d-1}} \left[\int_{\theta_i \in \mathbb{R}} g(\boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i^2} \pi(\boldsymbol{\theta}) \mathrm{d}\theta_i \right] \mathrm{d}\boldsymbol{\theta}_{-i},
\end{aligned}$$

and now integration by parts and using condition (17) and the Lipschitz continuity of g leads to

$$\begin{aligned}
&\int_{\theta_i \in \mathbb{R}} g(\boldsymbol{\theta}) \frac{\partial^2}{\partial \theta_i^2} \pi(\boldsymbol{\theta}) \mathrm{d}\theta_i \\
&= \left[-g(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} U_1(\boldsymbol{\theta}) \cdot \exp(-U_1(\boldsymbol{\theta})) \right]_{\theta_i = -\infty}^{\theta_i = \infty} \\
&+ \int_{\theta_i \in \mathbb{R}} \frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} U_1(\boldsymbol{\theta}) \exp(-U_1(\boldsymbol{\theta})) \mathrm{d}\theta_i \\
&= \int_{\theta_i \in \mathbb{R}} \frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} U_1(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\theta_i.
\end{aligned}$$

By summing up in i, we obtain that

$$\begin{aligned}
\mathcal{F}(\pi) &\leq \frac{1}{2} \sup_{g \in C^1(\mathbb{R}^d): \|\nabla g\|_\infty \leq 1} \sum_{i=1}^{d} \int_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{\partial}{\partial \theta_i} U_1(\boldsymbol{\theta}) \frac{\partial}{\partial \theta_i} g(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&\leq \frac{1}{2} \int_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\nabla U_1(\boldsymbol{\theta})\| \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.
\end{aligned}$$

Using the monotonicity property of $F(\pi_\rho)$, now the second bound of the theorem follows based on formula (41). The finiteness of this integral follows from assumption 17.

Now we are going to consider the m-strongly convex and M-smooth U_1 case. In such situations, it is straightforward to see that condition (17) holds with $a_1 = m \left\|\boldsymbol{\theta}^\star\right\|^2 / 2$, $a_2 = m/2$, $a_3 = 0$, $a_4 = M$, $\alpha = 2$ and $\beta = 1$; where $\boldsymbol{\theta}^\star$ is the minimum of U_1. For the integral of the norm of the gradient, we have by Jensen's inequality

$$\int_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\nabla U_1(\boldsymbol{\theta})\| \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leq \left(\int_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\nabla U_1(\boldsymbol{\theta})\|^2 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \right)^{1/2}.$$

For some index $1 \leq i \leq d$, we have

$$\int_{\boldsymbol{\theta} \in \mathbb{R}^d} \left(\frac{\partial}{\partial \theta_i} U_1(\boldsymbol{\theta}) \right)^2 \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \int_{\boldsymbol{\theta}_{-i} \in \mathbb{R}^{d-1}} \left[\int_{\theta_i \in \mathbb{R}} \left(\frac{\partial}{\partial \theta_i} U_1(\boldsymbol{\theta}) \right)^2 \exp\left(-U_1(\boldsymbol{\theta})\right) \mathrm{d}\theta_i \right] \mathrm{d}\boldsymbol{\theta}_{-i},$$

and using integration by parts, and the conditions of strong convexity and smoothness, we have

$$\begin{aligned}
&\int_{\theta_i\in\mathbb{R}} \left(\frac{\partial}{\partial\theta_i}U_1(\boldsymbol{\theta})\right)^2 \exp\left(-U_1(\boldsymbol{\theta})\right) \mathrm{d}\theta_i \\
&= \left[-\exp\left(-U_1(\boldsymbol{\theta})\right) \frac{\partial}{\partial\theta_i}U_1(\boldsymbol{\theta})\right]_{\theta_i=-\infty}^{\theta_i=\infty} + \int_{\theta_i\in\mathbb{R}} \exp\left(-U_1(\boldsymbol{\theta})\right) \frac{\partial^2}{\partial\theta_i^2}U_1(\boldsymbol{\theta})\mathrm{d}\theta_i \\
&\leq \int_{\theta_i\in\mathbb{R}} \exp\left(-U_1(\boldsymbol{\theta})\right) M_1 \mathrm{d}\theta_i.
\end{aligned}$$

By integrating this expression w.r.t. $\boldsymbol{\theta}_{-i}$ and summing up in i, we obtain that

$$\int_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\nabla U_1(\boldsymbol{\theta})\|^2 \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq M_1 d,$$

so the last claim of the theorem follows. ■

Appendix Appendix C. Proofs of the Results of Section 4

This section aims at proving the results claimed in Section 4.

Appendix C.1. Proof of Theorem 11

The following two propositions are going to be used for the proof of Theorem 11. The first one will allow us to bound the Wasserstein distance of two log-concave distributions based on the differences between their gradients. This is achieved by coupling processes evolving according to the Langevin dynamics with common Brownian noise.

Proposition 22 *Let μ and μ' be two distributions on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure, and whose negative log-likelihoods are continuously differentiable, strongly convex and smooth (gradient-Lipschitz). Denote the strong convexity constants $m(\mu), m(\mu')$ and smoothness constants $M(\mu)$ and $M(\mu')$. Then the Wasserstein distance of order $1 \leq p \leq \infty$ of these two distributions can be upper bounded as*

$$W_p(\mu,\mu') \leq \frac{\|D_{\mu,\mu'}\|_{L^p(\mu)}}{m(\mu')} \quad \text{for} \quad D_{\mu,\mu'}(\mathbf{z}) = \nabla \log \mu(\mathbf{z}) - \nabla \log \mu'(\mathbf{z}).$$

Proof Let $\mu(\mathbf{z}) = \exp(-U(\mathbf{z}))$ and $\mu'(\mathbf{z}) = \exp(-U'(\mathbf{z}))$.

First, we are going to consider the case $1 \leq p < \infty$. Note that it is easy to show that under the strong convexity and smoothness assumptions of this proposition, the Wasserstein distance of order p between μ and μ' is finite for such p. Assume that $(\mathbf{X}_1(0), \mathbf{X}_3(0))$ is an optimal coupling in Wasserstein distance of order p between μ and μ', so that $\mathbf{X}_1(0) \sim \mu$, $\mathbf{X}_3(0) \sim \mu'$, and

$$\left[\mathbb{E}\left(\|\mathbf{X}_1(0) - \mathbf{X}_3(0)\|^p\right)\right]^{1/p} = W_p(\mu,\mu').$$

The existence of such a coupling follows from Theorem 4.1 of Villani (2008). Let $\mathbf{X}_2(0) = \mathbf{X}_1(0)$. We now define three Langevin diffusions $(\mathbf{X}_1(t), \mathbf{X}_2(t), \mathbf{X}_3(t))_{t\geq 0}$ with a common noise (synchronous coupling)

$$\begin{aligned}
\mathrm{d}\mathbf{X}_1(t) &= -\nabla U(\mathbf{X}_1(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \\
\mathrm{d}\mathbf{X}_2(t) &= -\nabla U'(\mathbf{X}_2(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \\
\mathrm{d}\mathbf{X}_3(t) &= -\nabla U'(\mathbf{X}_3(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t.
\end{aligned}$$

Under the strong convexity and smoothness assumptions on the log-densities, these SDEs admit unique strong solutions (see Theorem 3.1 of Pavliotis (2014) and Arnold (1974)). Since $\mathbf{X_1}(0) \sim \mu$ and $X_3 \sim \mu'$, we can see that $\mathbf{X}_1(t) \sim \mu$ and $\mathbf{X}_3(t) \sim \mu'$ for every $t \geq 0$. $\mathbf{X}_2(t)$ is initialized at μ since $\mathbf{X}_2(0) = \mathbf{X}_1(0)$ and converges towards μ'. The proof of this proposition is based on a coupling argument based on these three diffusions. Let

$$D_{12}(t) = \mathbf{X}_1(t) - \mathbf{X}_2(t) - t(\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))).$$

Then we can decompose $\mathbf{X}_1(t) - \mathbf{X}_3(t)$ as

$$\mathbf{X}_1(t) - \mathbf{X}_3(t) = t(\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))) + D_{12}(t) + (\mathbf{X}_2(t) - \mathbf{X}_3(t)). \tag{42}$$

In the next two paragraphs of the proof, we are going to establish the following auxiliary inequalities

$$\|\mathbf{X}_2(t) - \mathbf{X}_3(t)\| \leq \exp(-m(\mu')t) \cdot \|\mathbf{X}_1(0) - \mathbf{X}_3(0)\|, \tag{43}$$

$$\|D_{12}(t)\| \leq C_0 t^3 + C_1 t^2(\|\nabla U(\mathbf{X}_1(0))\| + \|\nabla U'(\mathbf{X}_1(0))\|) + C_2 t \sup_{0\leq s\leq t} \|\mathbf{B}_s\| \text{ for } 0 \leq t \leq C_3, \tag{44}$$

for positive constants C_0, C_1, C_2, C_3 that only depend on the dimension d and the convexity parameters $m(\mu), m(\mu'), M(\mu), M(\mu')$. Let $\|X\|_{L^p} = (\mathbb{E}(\|X\|^p))^{1/p}$ denote the L^p norm of a random variable. By taking the L^p norms of both sides of (42), and using Minkowski's inequality, we can see that

$$\|\mathbf{X}_1(t) - \mathbf{X}_3(t)\|_{L^p} \leq t\|\nabla U'(\mathbf{X}_2(0)) - \nabla U(\mathbf{X}_1(0))\|_{L^p} + \|D_{12}(t)\|_{L^p} + \|\mathbf{X}_2(t) - \mathbf{X}_3(t)\|_{L^p}.$$

By the definition of the Wasserstein distance, we know that $W_p(\mu, \mu') \leq \|\mathbf{X}_1(t) - \mathbf{X}_3(t)\|_{L^p}$, and by assuming inequalities (43) and (44) are true, we obtain that for $0 \leq t \leq C_3$,

$$\begin{aligned}
W_p(\mu, \mu') \leq\ & t\|\nabla U'(\mathbf{X}_1(0)) - \nabla U(\mathbf{X}_1(0))\|_{L^p} + W_p(\mu, \mu')\exp(-m(\mu')t) \\
& + C_0 t^3 + C_1 t^2(\|\nabla U(\mathbf{X}_1(0))\|_{L^p} + \|\nabla U'(\mathbf{X}_2(0))\|_{L^p}) + C_2 t \left\| \sup_{0\leq s\leq t} \|\mathbf{B}_s\| \right\|_{L^p}.
\end{aligned} \tag{45}$$

It is easy to show that under the strong convexity and smoothness assumptions of this proposition, the terms $\|\nabla U(\mathbf{X}_1(0))\|_{L^p}$ and $\|\nabla U'(\mathbf{X}_1(0))\|_{L^p}$ are finite. By the reflection principle for the Brownian motion (see Lévy (1940)), in one dimension, the distribution of $\sup_{0\leq s\leq t} B_s$ is the same as the distribution of $|B_t|$. Using the triangle inequality, and the fact that $\|Y\|_{L^p} \leq \sqrt{p}$ for a standard Gaussian random variable Y, it follows that

$$\left\| \sup_{0\leq s\leq t} \|\mathbf{B}_s\| \right\|_{L^p} \leq 2d\sqrt{t}\sqrt{p}.$$

Hence all of the terms bounding $\|D_{12}(t)\|_{L^p}$ in (45) are of order $o(t)$, and the claim of the proposition follows by rearrangement and letting $t \searrow 0$.

Now we are going to prove the two auxiliary inequalities. We start with (43). From Itô's formula (see Lemma 3.2 of Pavliotis (2014)), $\|\mathbf{X}_2(t)-\mathbf{X}_3(t)\|^2$ is differentiable in t and satisfies

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{X}_2(t)-\mathbf{X}_3(t)\|^2 &= -2\left\langle \mathbf{X}_2(t)-\mathbf{X}_3(t), \nabla U'(\mathbf{X}_2(t)-\nabla U'(\mathbf{X}_3(t))\right\rangle \\ &\le -2m(\mu')\|\mathbf{X}_2(t)-\mathbf{X}_3(t)\|^2,\end{aligned}$$

where the last step follows from the strong convexity of U'. We obtain (43) by Grönwall's inequality and rearrangement.

We continue with the proof of (44). By Itô's formula, we can see that

$$\begin{aligned}D_{12}(t) &= \mathbf{X}_1(t)-\mathbf{X}_2(t)-t(\nabla U'(\mathbf{X}_1(0))-\nabla U(\mathbf{X}_1(0))) \\ &= \int_{s=0}^t (\nabla U'(\mathbf{X}_2(t))-\nabla U(\mathbf{X}_1(t)))\mathrm{d}s - t(\nabla U'(\mathbf{X}_1(0))-\nabla U(\mathbf{X}_1(0))) \\ &= \int_{s=0}^t [\nabla U'(\mathbf{X}_2(t))-\nabla U'(\mathbf{X}_1(0))]\mathrm{d}s + \int_{s=0}^t [\nabla U(\mathbf{X}_1(0))-\nabla U(\mathbf{X}_1(t))]\mathrm{d}s.\end{aligned}$$

Using the smoothness assumption for U and U', and the fact that $\mathbf{X}_1(0)=\mathbf{X}_2(0)$, we have

$$\|D_{12}(t)\| \le M(\mu')\int_{s=0}^t \|\mathbf{X}_2(t)-\mathbf{X}_2(0)\|\mathrm{d}s + M(\mu)\int_{s=0}^t \|\mathbf{X}_1(t)-\mathbf{X}_1(0)\|\mathrm{d}s. \tag{46}$$

Let

$$\begin{aligned}\mathbf{Y}_1'(t) &= -\nabla U(\mathbf{Y}_1(t)), \\ \mathbf{Y}_2'(t) &= -\nabla U'(\mathbf{Y}_2(t)),\end{aligned}$$

and assume that $\mathbf{Y}_1(0)=\mathbf{Y}_2(0)=\mathbf{X}_1(0)=\mathbf{X}_2(0)$. Then these ODEs have a unique solution (see page 74 of Perko (2013)). Now by the triangle inequality, and the fact that $\mathbf{Y}_1(0)=\mathbf{X}_1(0)$, we have

$$\|\mathbf{X}_1(s)-\mathbf{X}_1(0)\| \le \|\mathbf{Y}_1(s)-\mathbf{Y}_1(0)\| + \|\mathbf{Y}_1(s)-\mathbf{X}_1(s)\|.$$

For the first part, by Taylor's expansion, and the smoothness assumption on U, we have

$$\|\mathbf{Y}_1(t)-\mathbf{Y}_1(0)\| \le s\|\nabla U(\mathbf{Y}_1(0))\| + \frac{1}{2}M(\mu)s^2.$$

For the second part, by Itô's formula, we have

$$\begin{aligned}\mathbf{Y}_1(s)-\mathbf{X}_1(s) &= \int_{r=0}^s [\nabla U(\mathbf{X}_1(r))-\nabla U(\mathbf{Y}_1(r))]\mathrm{d}r + \sqrt{2}\mathbf{B}_s, \\ \|\mathbf{Y}_1(s)-\mathbf{X}_1(s)\| &\le M(\mu)\int_{r=0}^s \|\mathbf{X}_1(r)-\mathbf{Y}_1(r)\|\mathrm{d}r + \sqrt{2}\|\mathbf{B}_s\|, \\ \sup_{0\le r\le s}\|\mathbf{Y}_1(s)-\mathbf{X}_1(s)\| &\le M(\mu)s\sup_{0\le r\le s}\|\mathbf{Y}_1(s)-\mathbf{X}_1(s)\| + \sqrt{2}\sup_{0\le r\le s}\|\mathbf{B}_r\|.\end{aligned}$$

Hence for $s \leq 1/(2M(\mu))$, we have $\sup_{0\leq r\leq s} \|\mathbf{Y}_1(s) - \mathbf{X}_1(s)\| \leq 2\sqrt{2}\sup_{0\leq r\leq s} \|\mathbf{B}_r\|$. By combining the above two bounds, for $0 \leq s \leq 1/(2M(\mu))$, we have

$$\|\mathbf{X}_1(s) - \mathbf{X}_1(0)\| \leq s\|\nabla U(\mathbf{Y}_1(0))\| + \frac{1}{2}M(\mu)s^2 + 2\sqrt{2}\sup_{0\leq r\leq s}\|\mathbf{B}_r\|,$$

and by the same argument, for $0 \leq s \leq 1/(2M(\mu'))$,

$$\|\mathbf{X}_2(s) - \mathbf{X}_2(0)\| \leq s\|\nabla U'(W_2(0))\| + \frac{1}{2}M(\mu')s^2 + 2\sqrt{2}\sup_{0\leq r\leq s}\|\mathbf{B}_r\|.$$

Inequality now (44) follows by substituting these into (46) and doing some rearrangement.

Finally, the result for $p = \infty$ follows from a limiting argument. By Proposition 3 of Givens and Shortt (1984), we have

$$W_\infty(\mu,\mu') = \lim_{p\to\infty} W_p(\mu,\mu') \leq \sup_{1\leq p<\infty} \frac{\|D_{\mu,\mu'}\|_{L^p(\mu)}}{m(\mu')} \leq \frac{\|D_{\mu,\mu'}\|_{L^\infty(\mu)}}{m(\mu')}.$$

■

Proposition 23 *Let $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ be two parameter values, and μ_i, resp. μ_i', denotes the conditional distributions of $\mathbf{z}_i$ given $\boldsymbol{\theta}$ under Π_ρ, resp. $\boldsymbol{\theta}'$. Then under Assumption (A_4), for every $1 \leq p \leq \infty$, we have*

$$W_p(\mu_i,\mu_i') \leq \frac{1}{1+\rho^2 m_i}\|\mathbf{A}_i(\boldsymbol{\theta} - \boldsymbol{\theta}')\|. \tag{47}$$

Proof We have $\mu_i(\mathbf{z}) \propto \exp\left(-U_i(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta}-\mathbf{z}\|^2}{2\rho^2}\right)$ and $\mu_i'(\mathbf{z}) \propto \exp\left(-U_i(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta}'-\mathbf{z}\|^2}{2\rho^2}\right)$.

Proposition 22 requires the smoothness (gradient Lipschitz) property, so it cannot be applied directly to these potentials under our assumptions. To overcome this difficulty, we are going to use the Moreau-Yosida envelope of U_i (Durmus et al., 2018) defined for any regularisation parameter $\lambda > 0$ as

$$U_i^\lambda(\mathbf{z}) := \min_{\mathbf{y}\in\mathbb{R}^d}\left\{U_i(\mathbf{y}) + (2\lambda)^{-1}\|\mathbf{y}-\mathbf{z}\|^2\right\}.$$

By Theorem 1.25 of Rockafellar and Wets (1998), U_i^λ converges pointwise to U_i, that is for any $\mathbf{z} \in \mathbb{R}^d$,

$$\lim_{\lambda\to 0} U_i^\lambda(\mathbf{z}) = U_i(\mathbf{z}). \tag{48}$$

Moreover, from Proposition 12.19 of Rockafellar and Wets (1998) and Theorem 2.2 of Lemaréchal and Sagastizábal (1997) it follows that U_i^λ is λ^{-1} gradient Lipschitz and $\frac{m_i}{1+\lambda m_i}$-strongly convex.

Let $\mu_i^\lambda(\mathbf{z}) \propto \exp\left(-U_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta}-\mathbf{z}\|^2}{2\rho^2}\right)$ and $\mu_i'^\lambda(\mathbf{z}) \propto \exp\left(-U_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{A}_i\boldsymbol{\theta}'-\mathbf{z}\|^2}{2\rho^2}\right)$, then we have

$$\|\nabla \log(\mu_i^\lambda(\mathbf{z})) - \nabla \log(\mu_i'^\lambda(\mathbf{z}))\| = \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{A}_i\boldsymbol{\theta}'\|}{\rho^2}.$$

Since $-\log \mu_i^\lambda(\mathbf{z})$ and $-\log \mu_i'^\lambda(\mathbf{z})$ are $\frac{m_i}{1+\lambda m_i} + \frac{1}{\rho^2}$-strongly convex and $\frac{1}{\lambda} + \frac{1}{\rho^2}$-smooth, it follows from Proposition 22 that we have for every $1 \leq p \leq \infty$

$$W_p(\mu_i^\lambda, \mu_i'^\lambda) \leq \frac{\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{A}_i\boldsymbol{\theta}'\|}{1 + \rho^2 m_i/(1+m_i\lambda)}. \tag{49}$$

Now we are going to consider the case $1 \leq p < \infty$ first. To complete the proof, we still need to bound $W_p(\mu_i^\lambda, \mu_i)$. By Theorem 6.15 of Villani (2008), we have

$$W_p(\mu_i^\lambda, \mu_i) \leq \left[\int_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z} - \boldsymbol{\theta}\|^p |\mu_i(\mathbf{z}) - \mu_i^\lambda(\mathbf{z})| \mathrm{d}\mathbf{z}\right]^{1/p}. \tag{50}$$

Note that $|\mu_i(\mathbf{z}) - \mu_i^\lambda(\mathbf{z})| \leq \mu_i(\mathbf{z}) + \mu_i^\lambda(\mathbf{z})$. Moreover, from the definition of the Moreau-Yosida envelope U_i^λ, it follows that $U_i^\lambda(\mathbf{z}) \leq U_i^{\lambda'}(\mathbf{z})$ for $\lambda' < \lambda$, hence it is monotone increasing towards $U_i(\mathbf{z})$ as $\lambda \to 0$. This implies that the normalising constant

$$Z_i^\lambda = \int_{\mathbf{z}} \exp\left(-U_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{z} - \mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right) \mathrm{d}\mathbf{z}$$

is monotone decreasing towards $Z_i = \int_{\mathbf{z}} \exp\left(-U_i(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right) \mathrm{d}\mathbf{z}$ as $\lambda \to 0$ by the monotone convergence theorem. Therefore we have for any fixed $\Lambda > 0$ and $0 < \lambda < \Lambda$

$$\mu_i^\lambda(\mathbf{z}) = \frac{\exp\left(-U_i^\lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i^\lambda} \leq \frac{\exp\left(-U_i^\Lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i}.$$

This means that for $\lambda < \Lambda$, we have

$$\|\mathbf{z} - \boldsymbol{\theta}\|^p |\mu_i(\mathbf{z}) - \mu_i^\lambda(\mathbf{z})| \leq \|\mathbf{z} - \boldsymbol{\theta}\|^p \left(\mu_i(\mathbf{z}) + \frac{\exp\left(-U_i^\Lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i}\right).$$

Using the strong-convexity of $-\log \mu_i$, it follows that it has a unique minimizer which we denote by $\mathbf{z}_i^*$. In particular, we have

$$\int_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z} - \boldsymbol{\theta}\|^p \mu_i(\mathbf{z}) \mathrm{d}\mathbf{z} \leq \mu_i(\mathbf{z}_i^*) \int_{z\in\mathbb{R}^d} \|\mathbf{z} - \boldsymbol{\theta}\|^p \exp\left(-(m_i + 1/\rho^2)\|\mathbf{z} - \mathbf{z}_i^*\|^2/2\right) \mathrm{d}\mathbf{z} < \infty,$$

and with the same argument we can also show that

$$\int_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{z} - \boldsymbol{\theta}\|^p \frac{\exp\left(-U_i^\Lambda(\mathbf{z}) - \frac{\|\mathbf{z}-\mathbf{A}_i\boldsymbol{\theta}\|^2}{2\rho^2}\right)}{Z_i} < \infty.$$

Hence using the pointwise convergence (48) it follows from the dominated convergence theorem and the bound (50) that $W_p(\mu_i^\lambda, \mu_i) \to 0$ as $\lambda \to 0$. The same also holds for $W_p({\mu'_i}^\lambda, \mu'_i)$, so we can conclude using (49) and the triangle inequality

$$W_p(\mu_i, \mu'_i) \leq W_p(\mu_i, \mu_i^\lambda) + W_p({\mu_i}^\lambda, {\mu'_i}^\lambda) + W_p({\mu'_i}^\lambda, \mu'_i).$$

Finally, since we have shown the inequality (47) for $1 \leq p < \infty$, the bound for $p = \infty$ follows by Proposition 3 of Givens and Shortt (1984). ■

The following result is an elementary fact from linear algebra (proof is included for completeness).

Lemma 24 *Suppose that $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, and $\|\mathbf{v}\| \leq \|\mathbf{u}\|$. Then there exists a symmetric matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ such that $\mathbf{W}\mathbf{u} = \mathbf{v}$, and $-\mathbf{I} \preceq \mathbf{W} \preceq \mathbf{I}$ ($\preceq$ denotes the partial Loewner ordering).*

Proof First we assume that $\|\mathbf{u}\| = \|\mathbf{v}\|$. If $\mathbf{u} = \mathbf{v}$, then $\mathbf{W} = \mathbf{I}$ works, otherwise it is easy to check that

$$\mathbf{W} = (\mathbf{u} + \mathbf{v})(\mathbf{u} + \mathbf{v})^\top / \|\mathbf{u} + \mathbf{v}\|^2 - (\mathbf{u} - \mathbf{v})(\mathbf{u} - \mathbf{v})^\top / \|\mathbf{u} - \mathbf{v}\|^2$$

satisfies the requirements. The general case follows by rescaling. ■

Now we are ready to prove our contraction bound.
Proof [Proof of Theorem 11] Let $(\mathbf{Z}_{1:b}, \mathbf{Z}'_{1:b})$ be a coupling of the two distributions $\Pi_\rho(\mathbf{Z}_{1:b}|\boldsymbol{\theta})$ and $\Pi_\rho(\mathbf{Z}'_{1:b}|\boldsymbol{\theta})$ such that

$$\|\mathbf{Z}_i - \mathbf{Z}'_i\| \leq \frac{1}{1 + \rho^2 m_i} \|\mathbf{A}_i(\boldsymbol{\theta} - \boldsymbol{\theta}')\| \text{ almost surely.} \tag{51}$$

The existence of such a coupling follows from Proposition 23. Given this coupling $(\mathbf{Z}_{1:b}, \mathbf{Z}'_{1:b})$, our next step is to couple the two conditional distributions

$$\Pi_\rho(\boldsymbol{\theta}|\mathbf{Z}_{1:b}) \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}(\mathbf{Z}_{1:b}), \Sigma_{\boldsymbol{\theta}}),$$
$$\Pi_\rho(\boldsymbol{\theta}|\mathbf{Z}'_{1:b}) \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}(\mathbf{Z}'_{1:b}), \Sigma_{\boldsymbol{\theta}}),$$

where $\Sigma_{\boldsymbol{\theta}} = \rho^2 (\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i)^{-1}$ and $\mu_{b\theta}(\mathbf{z}_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i)^{-1} \sum_{i=1}^b \mathbf{A}_i^\top \mathbf{z}_i$. Since these two Gaussian distributions have the same covariance matrix, coupling them can be done in a straightforward way, and we can see that for the metric w introduced in the statement of Theorem 11, for every $1 \leq p \leq \infty$, we have

$$W_p^w(\mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}, \cdot), \mathbf{P}_{\mathrm{SGS}}(\boldsymbol{\theta}', \cdot)) \leq \left[\mathbb{E}(w(\mu_{\boldsymbol{\theta}}(\mathbf{Z}_{1:b}), \mu_{\boldsymbol{\theta}}(\mathbf{Z}'_{1:b}))^p)\right]^{1/p}, \tag{52}$$

where W_p^w denotes Wasserstein distance of order p with respect to the metric w. Note that

$$\mu_{\boldsymbol{\theta}}(\mathbf{Z}_{1:b}) - \mu_{\boldsymbol{\theta}}(\mathbf{Z}'_{1:b}) = (\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i)^{-1} \sum_{i=1}^b \mathbf{A}_i^\top (\mathbf{Z}_i - \mathbf{Z}'_i).$$

For each $i \in [b]$, we now apply Lemma 24 with $\mathbf{v} = \mathbf{Z}_i - \mathbf{Z}'_i$ and $\mathbf{u} = \mathbf{A}_i(\boldsymbol{\theta} - \boldsymbol{\theta}')/(1+\rho^2 m_i)$. Using (51), the assumption $\|\mathbf{v}\| \leq \|\mathbf{u}\|$ of Lemma 24 is satisfied and there exist some symmetric matrices $\mathbf{W}_1, \ldots, \mathbf{W}_b \in \mathbb{R}^{d\times d}$ such that $-\mathbf{I} \preceq \mathbf{W}_i \preceq \mathbf{I}$, and

$$\mathbf{Z}_i - \mathbf{Z}'_i = \mathbf{W}_i \frac{\mathbf{A}_i(\boldsymbol{\theta} - \boldsymbol{\theta}')}{1+\rho^2 m_i}, \quad \forall i \in [b].$$

This yields

$$\mu_{\boldsymbol{\theta}}(\mathbf{Z}_{1:b}) - \mu_{\boldsymbol{\theta}}(\mathbf{Z}'_{1:b}) = \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1} \sum_{i=1}^b \frac{\mathbf{A}_i^\top \mathbf{W}_i \mathbf{A}_i}{1+\rho^2 m_i}(\boldsymbol{\theta} - \boldsymbol{\theta}') \text{ almost surely.}$$

From the definition of w, we can now write

$$\begin{aligned}
w(\mu_{\boldsymbol{\theta}}(\mathbf{Z}_{1:b}), \mu_{\boldsymbol{\theta}}(\mathbf{Z}'_{1:b})) &= \left\| \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{1/2} (\mu_{\boldsymbol{\theta}}(\mathbf{Z}_{1:b}) - \mu_{\boldsymbol{\theta}}(\mathbf{Z}'_{1:b})) \right\| \\
&= \left\| \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2} \sum_{i=1}^b \frac{\mathbf{A}_i^\top \mathbf{W}_i \mathbf{A}_i}{1+\rho^2 m_i}(\boldsymbol{\theta} - \boldsymbol{\theta}') \right\| \\
&= \left\| \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2} \sum_{i=1}^b \frac{\mathbf{A}_i^\top \mathbf{W}_i \mathbf{A}_i}{1+\rho^2 m_i} \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2} \cdot \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}') \right\| \\
&\leq \left\| \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2} \left(\sum_{i=1}^b \frac{\mathbf{A}_i^\top \mathbf{A}_i}{1+\rho^2 m_i}\right) \left(\sum_{i=1}^b \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2} \right\| w(\boldsymbol{\theta}, \boldsymbol{\theta}') \text{ almost surely.}
\end{aligned}$$

Hence the result follows from (52). ■

Appendix C.2. Proof of Corollary 12

First, we will show the convergence results in Wasserstein distance of order p for $1 \leq p < \infty$. Let $(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)$ be the optimal coupling of the initial distribution ν and the stationary distribution π_ρ that achieves the Wasserstein distance of order p for the metric w (see Theorem 4.1 of Villani (2008) for proof of existence), i.e.

$$W_p^w(\nu, \pi_\rho) = \|w(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)\|_{L^p}.$$

For $i \geq 1$, assuming that $(\boldsymbol{\theta}_{0:i-1}, \boldsymbol{\theta}'_{0:i-1})$ has been defined, add two more elements $(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i)$ by defining their conditional distribution based on the past elements as the optimal coupling between $P_{\mathrm{SGS}}(\boldsymbol{\theta}_{i-1}, \cdot)$ and $P_{\mathrm{SGS}}(\boldsymbol{\theta}'_{i-1}, \cdot)$ achieving the Wasserstein distance of order p for the metric w. Using that $K_p(\boldsymbol{\theta}, \boldsymbol{\theta}') \geq K_{\mathrm{SGS}}$ by Theorem 11, we have

$$\mathbb{E}(w(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_1)^p | \boldsymbol{\theta}_0, \boldsymbol{\theta}'_0) \leq (1 - K_{\mathrm{SGS}})^p w(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)^p,$$

and so by the tower property, we have

$$\|w(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_1)\|_{L^p} \leq (1 - K_{\mathrm{SGS}}) W_p^w(\nu, \pi_\rho).$$

Similarly, by induction, it follows that

$$\|w(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i)\|_{L^p} \leq (1 - K_{\mathrm{SGS}})^i W_p^w(\nu, \pi_\rho).$$

Now (20) for $1 \leq p < \infty$ follows by noticing that $\boldsymbol{\theta}'_i \sim \pi_\rho$ since the Markov chain $(\boldsymbol{\theta}'_j)_{j \geq 0}$ was initialized in its stationary distribution. Finally, the $p = \infty$ case follows from Proposition 3 of Givens and Shortt (1984).

Regarding the convergence rate in total variation distance stated in Theorem 11, we will use Corollary 25 and Proposition 26 detailed below.

Corollary 25 (Lower bound on the spectral gap of SGS) *SGS defines a reversible Markov chain. Under Assumptions* (A_4) *and* (A_5)*, its absolute spectral gap* γ^*_{SGS} *is lower bounded by* K_{SGS}*, see* (19)*.*

Proof The reversibility follows by a standard argument for data augmentation schemes given in Lemma 3.1 of Liu et al. (1994). The lower bound on the absolute spectral gap follows by Proposition 30 of Ollivier (2009). ■

The following proposition is well known in the MCMC literature but we have only found a proof for Markov chains on finite state spaces. Hence for completeness, we include a short proof here.

Proposition 26 *Suppose that* $\mathbf{P}(\mathbf{z}, \cdot)$ *is a reversible Markov kernel on a Polish state space* Ω *with absolute spectral gap* $\gamma^* > 0$*, and unique stationary distribution* π*. Then for any initial distribution* ν *that is absolutely continuous with respect to* π*, and any number of steps* $t \in \mathbb{Z}_+$*, we have*

$$\left\| \nu \mathbf{P}^t - \pi \right\|_{\mathrm{TV}} \leq \frac{1}{2} \left(\mathbb{E}_\pi \left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right)^2 \right] - 1 \right)^{1/2} \cdot (1 - \gamma^*)^t.$$

Our proof is based on the following lemma.

Lemma 27 *Suppose that* $\mathbf{Q}(x, \mathrm{d}y)$ *is a reversible Markov kernel on a Polish state space* Ω *with stationary distribution* π*. Then for any distribution* ν *that is absolutely continuous with respect to* π*,* $\nu\mathbf{Q}$ *is also absolutely continuous with respect to* π*, and for* π*-almost every* $x \in \Omega$*, we have*

$$\frac{\mathrm{d}(\nu\mathbf{Q})}{\mathrm{d}\pi}(x) = \left(\mathbf{Q} \left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi} \right) \right)(x).$$

Proof The claim of the lemma is equivalent to showing that for every bounded measurable function $f : \Omega \to \mathbb{R}$, we have

$$\int_{x\in\Omega} \frac{\mathrm{d}(\nu\mathbf{Q})}{\mathrm{d}\pi}(x) f(x)\pi(\mathrm{d}x) = \int_{x\in\Omega} \left(\mathbf{Q}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x) f(x)\pi(\mathrm{d}x). \tag{53}$$

Since if we add a constant to f, both sides increase by this constant, we can assume without loss of generality that f is non-negative. Under this assumption, we have

$$\begin{aligned}
&\int_{x\in\Omega} \frac{\mathrm{d}(\nu\mathbf{Q})}{\mathrm{d}\pi}(x) f(x)\pi(\mathrm{d}x) = \int_{x\in\Omega} f(x)(\nu\mathbf{Q})(\mathrm{d}x) \\
&= \int_{x,y\in\Omega} f(x)\nu(\mathrm{d}y)\mathbf{Q}(y,\mathrm{d}x) = \int_{x,y\in\Omega} f(y)\nu(\mathrm{d}x)\mathbf{Q}(x,\mathrm{d}y) \\
&= \int_{x,y\in\Omega} f(y)\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x)\pi(\mathrm{d}x)\mathbf{Q}(x,\mathrm{d}y)
\end{aligned}$$

by the monotone convergence theorem (using the non-negativity of f)

$$= \lim_{M\to\infty} \int_{x,y\in\Omega} f(y)\min\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x), M\right)\pi(\mathrm{d}x)\mathbf{Q}(x,\mathrm{d}y)$$

using the reversibility of $\mathbf{Q}$ (in the equivalent bounded measurable test function formulation)

$$= \lim_{M\to\infty} \int_{x,y\in\Omega} f(y)\min\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x), M\right)\pi(\mathrm{d}y)\mathbf{Q}(y,\mathrm{d}x)$$

by the monotone convergence theorem (using the non-negativity of f)

$$\begin{aligned}
&= \int_{x,y\in\Omega} f(y)\frac{\mathrm{d}\nu}{\mathrm{d}\pi}(x)\pi(\mathrm{d}y)\mathbf{Q}(y,\mathrm{d}x) \\
&= \int_{y\in\Omega} f(y)\left(\mathbf{Q}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(y)\pi(\mathrm{d}y),
\end{aligned}$$

hence (53) and the claim of our lemma holds. ■

Proof [Proof of Proposition 26] We define the Hilbert space $L^2(\pi)$ as measurable functions f on Ω satisfying $\mathbb{E}_\pi(f^2) < \infty$, endowed with the scalar product $\langle f, g\rangle_\pi = \int_{\mathbf{z}\in\Omega} f(\mathbf{z})g(\mathbf{z})\pi(\mathrm{d}\mathbf{z})$. Let us define the linear operator $\mathbf{\Pi}(f)(\mathbf{z}) := \mathbb{E}_\pi(f)$ for any $f \in L^2(\pi)$, $\mathbf{z} \in \Omega$.

Using Lemma 27 with $\mathbf{Q} = \mathbf{P}^t$, it follows that

$$\|\nu\mathbf{P}^t - \pi\|_{\mathrm{TV}} = \frac{1}{2}\int_{x\in\Omega} \left|\frac{\mathrm{d}\nu\mathbf{P}^t}{\mathrm{d}\pi}(x) - 1\right| \pi(\mathrm{d}x)$$

using Jensen's inequality, we have

$$\leq \frac{1}{2}\sqrt{\int_{x\in\Omega} \left(\frac{\mathrm{d}\nu\mathbf{P}^t}{\mathrm{d}\pi}(x) - 1\right)^2 \pi(\mathrm{d}x)}. \tag{54}$$

Using Lemma 27 again, the integral inside the square root can be further bounded as

$$\int_{x\in\Omega}\left(\frac{\mathrm{d}\nu\mathbf{P}^t}{\mathrm{d}\pi}(x)-1\right)^2\pi(\mathrm{d}x)=\int_{x\in\Omega}\left(\left(\mathbf{P}^t\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x)-1\right)^2\pi(\mathrm{d}x)$$
$$=\int_{x\in\Omega}\left(\left((\mathbf{P}^t-\mathbf{\Pi})\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x)\right)^2\pi(\mathrm{d}x)=\int_{x\in\Omega}\left(\left((\mathbf{P}-\mathbf{\Pi})^t\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right)\right)(x)\right)^2\pi(\mathrm{d}x)$$
$$=\int_{x\in\Omega}\left(\left((\mathbf{P}-\mathbf{\Pi})^t\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right)\right)(x)\right)^2\pi(\mathrm{d}x)=\left\langle\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1,(\mathbf{P}-\mathbf{\Pi})^{2t}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right)\right\rangle_\pi$$
$$\leq\|\mathbf{P}-\mathbf{\Pi}\|_\pi^{2t}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right\|_\pi^2=(1-\gamma^*)^{2t}\left\|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}-1\right\|_\pi^2,$$

and the claim of the proposition follows by substituting this into (54). ■

Now we are ready to prove our convergence bound in total variation distance.

Proof [Proof of Theorem 15] From Corollary 25, we know that the absolute spectral gap of SGS satisfies that $\gamma^*\geq K_{\mathrm{SGS}}$ (defined in (19)), and Proposition 26 implies that

$$\left\|\nu\mathbf{P}_{\mathrm{SGS}}^t-\pi_\rho\right\|_{\mathrm{TV}}\leq\sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right]-1}\cdot(1-\gamma^*)^t$$
$$\leq\sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right]-1}\cdot(1-K_{\mathrm{SGS}})^t.$$

■

Appendix C.3. Proof of Theorem 13

Proof [Proof of Theorem 13] From Theorem 10, it follows that if ρ is chosen as in (21), then

$$W_1(\pi_\rho,\pi)\leq\frac{\epsilon}{2}\cdot\frac{\sqrt{d}}{\sqrt{m_1}}. \tag{55}$$

From Proposition 1 part (ii) in Durmus and Moulines (2019) it follows that for the initial distribution $\delta_{\boldsymbol{\theta}^\star}$ (Dirac measure at $\boldsymbol{\theta}^\star$), we have

$$W_1(\delta_{\boldsymbol{\theta}^\star},\pi)\leq W_2(\delta_{\boldsymbol{\theta}^\star},\pi)\leq\frac{\sqrt{d}}{\sqrt{m_1}},$$

and hence by combining this with (55) using the triangle inequality and the assumption $\epsilon\leq 1$, it follows that

$$W_1(\delta_{\boldsymbol{\theta}^\star},\pi_\rho)\leq\frac{3}{2}\frac{\sqrt{d}}{\sqrt{m_1}}.$$

Now from Theorem 11, it follows that the coarse Ricci curvature of SGS is lower bounded by

$$K_{\mathrm{SGS}} := \frac{\rho^2 m_1}{1+\rho^2 m_1},$$

and therefore by Corollary 21 of Ollivier (2009), we have

$$W_1(P^t_{\mathrm{SGS}}(\boldsymbol{\theta}^\star, \cdot), \pi_\rho) \leq W_1(\delta_{\boldsymbol{\theta}^\star}, \pi_\rho) \cdot (1 - K_{\mathrm{SGS}})^t \leq \frac{\epsilon}{2} \cdot \frac{\sqrt{d}}{\sqrt{m_1}}.$$

The claim of the theorem now follows by the triangle inequality. ■

Appendix C.4. Complexity Bounds for Implementing SGS by Rejection Sampling

The following bound is a standard result in rejection sampling; see for instance Section 2.3 of Robert and Casella (2013).

Lemma 28 *Suppose that $\mu(\mathbf{z}) = \tilde{\mu}(\mathbf{z})/\tilde{Z}$ is the target density on $\mathbb{R}^d$, and $\nu(\mathbf{z})$ is the proposal density (both absolutely continuous w.r.t. the Lebesgue measure). Here $\tilde{\mu}(\mathbf{z})$ is the unnormalized target and $\tilde{Z}$ is the normalising constant (which is typically unknown). Suppose that the condition*

$$\tilde{\mu}(\mathbf{z}) \leq M\nu(\mathbf{z}) \tag{56}$$

holds for some constant $M < \infty$ for every $\mathbf{z} \in \mathbb{R}^d$. Under this assumption, if we take samples $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ from ν and accept $\mathbf{Z}_i$ with probability $\frac{\tilde{\mu}(\mathbf{Z}_i)}{M\nu(\mathbf{Z}_i)}$, then the accepted samples will be distributed according to μ. Moreover, the expected number of samples taken until the first acceptance is equal to $M/\tilde{Z}$.

The following lemma gives a complexity bound for rejection sampling for log-concave distributions. We assume that we have access to an approximation of the minimum of the strongly convex and smooth potential U, which will be denoted by $\tilde{\mathbf{z}}$. The quality of this approximation is taken into account in the proposal distribution using the norm of $\nabla U(\tilde{\mathbf{z}})$.

Lemma 29 (Rejection sampling upper bound for log-concave densities) *Suppose that $\mu(\mathbf{z}) \propto \exp(-U(\mathbf{z}))$ is a distribution on $\mathbb{R}^d$ such that U is twice differentiable and*

$$A\mathbf{I}_d \preceq \nabla^2 U(\mathbf{z}) \preceq B\mathbf{I}_d \tag{57}$$

for some $0 < A \leq B$ (strongly convex and smooth). Let $\mathbf{z}^*$ be the unique minimizer of U, $\tilde{\mathbf{z}}$ another point (an approximation of $\mathbf{z}^*$), and $\nu(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \tilde{\mathbf{z}}, \tilde{A}^{-1}\mathbf{I}_d)$, where

$$\tilde{A} = A + \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2d} - \sqrt{\frac{\|\nabla U(\tilde{\mathbf{z}})\|^4}{4d^2} + \frac{A\|\nabla U(\tilde{\mathbf{z}})\|^2}{d}}. \tag{58}$$

Suppose that we take samples $\mathbf{Z}_1, \mathbf{Z}_2, \ldots$ *from* ν*, and accept them with probability*

$$\mathbb{P}(\mathbf{Z}_j \text{ is accepted}) = \exp\left(-\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})} - [U(\mathbf{z}) - U(\tilde{\mathbf{z}})] + \frac{\tilde{A}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2}{2}\right).$$

Then these accepted samples are distributed according to μ*. Moreover, the expected number of samples taken until one is accepted is less than or equal to* $\left(B/\tilde{A}\right)^{d/2} \cdot \exp\left[\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2}\left(\frac{1}{A-\tilde{A}} - \frac{1}{B}\right)\right]$.

Proof The proposal density equals

$$\begin{aligned}\nu(\mathbf{z}) &= \mathcal{N}(\mathbf{z}; \tilde{\mathbf{z}}, \tilde{A}^{-1}\mathbf{I}_d) \\ &= \exp\left(-\frac{\tilde{A}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2}{2}\right) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2}.\end{aligned}$$

We define the unnormalized version of μ as

$$\tilde{\mu}(\mathbf{z}) = \exp(-[U(\mathbf{z}) - U(\tilde{\mathbf{z}})]) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2}.$$

Notice that

$$U(\mathbf{z}) - U(\tilde{\mathbf{z}}) = \left\langle \int_{t=0}^{1} \nabla U(\tilde{\mathbf{z}} + t(\mathbf{z}-\tilde{\mathbf{z}}))\mathrm{d}t, \mathbf{z}-\tilde{\mathbf{z}} \right\rangle.$$

By the intermediate value theorem, there is some $\mathbf{z}(t)$ such that

$$= \langle \nabla U(\tilde{\mathbf{z}}), \mathbf{z}-\tilde{\mathbf{z}} \rangle + \left\langle \mathbf{z}-\tilde{\mathbf{z}}, \left(\int_{t=0}^{1} t\nabla^2 U(\mathbf{z}(t))\mathrm{d}t\right)^{\top} (\mathbf{z}-\tilde{\mathbf{z}}) \right\rangle,$$

so using the assumption (57) it follows that

$$\geq -\|\nabla U(\tilde{\mathbf{z}})\|\|\mathbf{z}-\tilde{\mathbf{z}}\| + \frac{A}{2}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2.$$

Based on this, one gets

$$\frac{\tilde{\mu}(\mathbf{z})}{\nu(\mathbf{z})} \leq \exp\left(\|\nabla U(\tilde{\mathbf{z}})\| \cdot \|\mathbf{z}-\tilde{\mathbf{z}}\| - \frac{A-\tilde{A}}{2}\|\mathbf{z}-\tilde{\mathbf{z}}\|^2\right) \leq \exp\left(\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})}\right).$$

Hence we have $\tilde{\mu}(\mathbf{z}) \leq M\nu(\mathbf{z})$ for $M = \exp\left(\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A-\tilde{A})}\right)$.

For the normalising constant, we have

$$\tilde{Z} = \int_{\mathbf{z}\in\mathbb{R}^d} \tilde{\mu}(\mathbf{z})\mathrm{d}\mathbf{z} = \exp(U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*)) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2} \cdot \int_{\mathbf{z}\in\mathbb{R}^d} \exp(-(U(\mathbf{z}) - U(\mathbf{z}^*)))\mathrm{d}\mathbf{z}$$

using Taylor's expansion with second order remainder term, and assumption (57) yields

$$\geq \exp(U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*)) \cdot \left(\frac{\tilde{A}}{2\pi}\right)^{d/2} \cdot \int_{\mathbf{z}\in\mathbb{R}^d} \exp\left(-\frac{B}{2}\|\mathbf{z}-\mathbf{z}^*\|^2\right)\mathrm{d}\mathbf{z}$$

$$= \left(\frac{\tilde{A}}{B}\right)^{d/2} \cdot \exp\left(U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*)\right) \geq \left(\frac{\tilde{A}}{B}\right)^{d/2} \exp\left(\frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2B}\right),$$

where in the last step we have used the fact that for $\mathbf{z}' = \tilde{\mathbf{z}} - \frac{\|\nabla U(\tilde{\mathbf{z}})\|}{B}$, we have

$$\begin{aligned} &U(\tilde{\mathbf{z}}) - U(\mathbf{z}^*) \\ &\geq U(\tilde{\mathbf{z}}) - U(\mathbf{z}') = \left\langle \int_{t=0}^{1} \nabla U(\tilde{\mathbf{z}} + t(\mathbf{z}' - \tilde{\mathbf{z}}))\mathrm{d}t, \tilde{\mathbf{z}} - \mathbf{z}' \right\rangle \end{aligned}$$

using the fact that $\mathbf{z}^*$ is the minimum of U.

By the intermediate value theorem, there is some $\tilde{\mathbf{z}}(t) \in \mathbb{R}^d$ such that

$$\begin{aligned} &= \left\langle \nabla U(\tilde{\mathbf{z}}), \tilde{\mathbf{z}} - \mathbf{z}' \right\rangle + \left\langle \tilde{\mathbf{z}} - \mathbf{z}', \left(\int_{t=0}^{1} t\nabla^2 U(\tilde{\mathbf{z}}(t))\mathrm{d}t \right) \cdot (\tilde{\mathbf{z}} - \mathbf{z}') \right\rangle \\ &\geq \left\langle \nabla U(\tilde{\mathbf{z}}), \tilde{\mathbf{z}} - \mathbf{z}' \right\rangle - \frac{B}{2}\|\tilde{\mathbf{z}} - \mathbf{z}'\|^2 = \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2B}. \end{aligned}$$

Now it follows by Lemma 28 and the above bound on $\tilde{Z}$ that the expected number of samples until the first acceptance is less than or equal to

$$E(\tilde{A}) := \exp\left(\|\nabla U(\tilde{\mathbf{z}})\|^2 \left(\frac{1}{2(A - \tilde{A})} - \frac{1}{2B} \right) \right) \left(\frac{B}{\tilde{A}} \right)^{d/2}.$$

The parameter $\tilde{A}$ in (58) is chosen such that $E(\tilde{A})$ is minimized. Note that the minimizer of $E(\tilde{A})$ is the same as the minimizer of

$$\log(E(\tilde{A})) = \frac{d}{2}\log(B) - \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2B} + \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A - \tilde{A})} - \frac{d}{2}\log(\tilde{A}).$$

It is easy to check that this is a strictly convex function of $\tilde{A}$ on the interval $(0, A)$, and hence the unique minimum is taken at a point where the derivative is zero. This point, denoted by $\tilde{A}_{\min}$, thus satisfies

$$\left.\frac{\partial \log(E(\tilde{A}))}{\partial \tilde{A}}\right|_{\tilde{A}=\tilde{A}_{\min}} = \frac{\|\nabla U(\tilde{\mathbf{z}})\|^2}{2(A - \tilde{A})^2} - \frac{d}{2} \cdot \frac{1}{\tilde{A}} = 0.$$

Hence by rearrangement

$$\begin{aligned} &(\tilde{A} - A)^2 - (\|\nabla U(\tilde{\mathbf{z}})\|^2/d)\tilde{A} = 0 \\ &\tilde{A}^2 - (2A + \|\nabla U(\tilde{\mathbf{z}})\|^2/d)\tilde{A} + A^2 = 0 \\ &\tilde{A} = \frac{(2A + \|\nabla U(\tilde{\mathbf{z}})\|^2/d) \pm \sqrt{(2A + \|\nabla U(\tilde{\mathbf{z}})\|^2/d)^2 - 4A^2}}{2} \\ &\quad = A + \|\nabla U(\tilde{\mathbf{z}})\|^2/(2d) \pm \sqrt{\|\nabla U(\tilde{\mathbf{z}})\|^4/(4d^2) + A\|\nabla U(\tilde{\mathbf{z}})\|^2/d}. \end{aligned}$$

Only the solution with the $-$ sign falls in the interval $(0, A)$, hence it is the minimizer of $M/\tilde{Z}$. ■

Proof [Proof of Proposition 1] The fact that the accepted samples are distributed according to $\Pi_\rho(\mathbf{z}_i|\boldsymbol{\theta})$ and the formula (9) about the expected number of samples until acceptance follows from Lemma 29.

Let $G := \|\nabla V_i(\tilde{\mathbf{z}}_i(\boldsymbol{\theta}))\|$, then $\tilde{A}_i = 1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}$, and we have

$$\log(E_i) = \frac{d_i}{2}\log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) + \frac{G^2}{2}\left(\frac{1}{\sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i} - G^2/(2d_i)} - \frac{1}{1/\rho^2 + M_i}\right). \tag{59}$$

For the first part, notice that

$$\begin{aligned}
&\log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) \\
&= \log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i}\right) + \log\left(\frac{1/\rho^2 + m_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) \\
&= \log\left(1 + \frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i}\right) + \log\left(\frac{1}{1 + c - \sqrt{c^2 + 2c}}\right),
\end{aligned}$$

where $c = \frac{G^2/(2d_i)}{1/\rho^2 + m_i}$. Now using the fact that $\log(1+x) \leq x$ for $x > 0$, and that $\log\left(\frac{1}{1+c-\sqrt{c^2+2c}}\right) \leq \sqrt{2c}$ for $c \geq 0$, it follows that we have

$$\begin{aligned}
&\frac{d_i}{2}\log\left(\frac{1/\rho^2 + M_i}{1/\rho^2 + m_i + G^2/(2d_i) - \sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i}}\right) \\
&\leq \frac{d_i}{2}\left(\frac{\rho^2(M_i - m_i)}{1 + \rho^2 m_i} + \frac{G}{\sqrt{d_i(1/\rho^2 + m_i)}}\right).
\end{aligned}$$

For the second part (59),

$$\begin{aligned}
&\frac{G^2}{2}\left(\frac{1}{\sqrt{G^4/(4d_i^2) + G^2\left(1/\rho^2 + m_i\right)/d_i} - G^2/(2d_i)} - \frac{1}{1/\rho^2 + M_i}\right) \\
&= \frac{d_i}{\sqrt{1 + 4\left(1/\rho^2 + m_i\right)d_i/G^2} - 1} - \frac{G^2}{2}\cdot\frac{1}{1/\rho^2 + M_i}
\end{aligned}$$

using the fact that $\frac{1}{\sqrt{1+x}-1} \leq \frac{2}{\sqrt{x}}$ for $x \geq 2$, for $G \leq \sqrt{2d_i(1/\rho^2+m_i)}$, we have

$$\leq G \cdot \frac{\sqrt{d_i}}{\sqrt{1/\rho^2+m_i}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2+M_i}.$$

Hence by combining these terms, we obtain that for $G \leq \sqrt{2d_i(1/\rho^2+m_i)}$,

$$\log(E_i) \leq \frac{d_i}{2}\frac{\rho^2(M_i-m_i)}{1+\rho^2 m_i} + G \cdot \frac{3}{2} \cdot \frac{\sqrt{d_i}}{\sqrt{(1/\rho^2+m_i)}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2+M_i}$$

Under the first part of assumption (10), $\rho^2(2d_i(M_i-m_i)-m_i) \leq 1$, one can check that $\frac{d_i}{2}\frac{\rho^2(M_i-m_i)}{1+\rho^2 m_i} \leq \frac{1}{4}$. Using the second part of (10), $G \leq \frac{2}{7} \cdot \frac{\sqrt{1/\rho^2+m_i}}{\sqrt{d_i}}$, it follows that $G \cdot \frac{3}{2} \cdot \frac{\sqrt{d_i}}{\sqrt{(1/\rho^2+m_i)}} - \frac{G^2}{2} \cdot \frac{1}{1/\rho^2+M_i} \leq \log(2) - \frac{1}{4}$, so $\log(E_i) \leq \log(2)$ and our claim holds. ■

Appendix C.5. Proof of Theorems 14 and 15

The next two lemmas will be used for obtaining our total variation distance convergence rates.

Lemma 30 *Suppose that $U : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and M-gradient-Lipschitz. Then for every $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\|\nabla U(\mathbf{x})\|^2 \leq 2M(U(\mathbf{x}) - \inf_{\mathbf{x}\in\mathbb{R}^d} U(\mathbf{x})).$$

Proof Let $\mathbf{x}' = \mathbf{x} - \nabla U(\mathbf{x})/M$, then we have

$$\begin{aligned} U(\mathbf{x}) - U(\mathbf{x}') &= \int_{t=0}^{1} \left\langle \nabla U(\mathbf{x}+t(\mathbf{x}'-\mathbf{x})), \mathbf{x}-\mathbf{x}' \right\rangle \mathrm{d}t \\ &= \left\langle \nabla U(\mathbf{x}), \mathbf{x}-\mathbf{x}' \right\rangle + \int_{t=0}^{1} \left\langle \nabla U(\mathbf{x}+t(\mathbf{x}'-\mathbf{x})) - \nabla U(\mathbf{x}), \mathbf{x}-\mathbf{x}' \right\rangle \mathrm{d}t \end{aligned}$$

using the M-gradient Lipschitz property

$$\begin{aligned} &\geq \left\langle \nabla U(\mathbf{x}), \mathbf{x}-\mathbf{x}' \right\rangle - \int_{t=0}^{1} Mt\|\mathbf{x}-\mathbf{x}'\|^2 \mathrm{d}t \\ &\geq \frac{\|\nabla U(\mathbf{x})\|^2}{2M}, \end{aligned}$$

hence the result. ■

Lemma 31 *Suppose that Assumptions (A_0), (A_2), (A_4), (A_6) and $\det\left(\sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i\right) > 0$ hold. Let $\boldsymbol{\theta}^\star$ be the unique minimizer of $U(\boldsymbol{\theta}) = \sum_{i=1}^b U_i(\mathbf{A}_i\boldsymbol{\theta})$, and*

$$\nu(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, \left(\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1}\right).$$

If $b = 1$, $d = d_1$, and $\mathbf{A}_1$ is of full rank, then for any $\rho > 0$, we have $\frac{\nu(\boldsymbol{\theta})}{\pi_\rho(\boldsymbol{\theta})} \leq C_\rho$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$, where

$$C_\rho := (1 + \rho^2 M_1)^{d/2} \cdot \left(\frac{M_1}{m_1}\right)^{\frac{d}{2}}. \tag{60}$$

More generally, for multiple splitting, for $\rho^2 \leq \frac{1}{6\sigma_U^2}$, we have $\frac{\nu(\boldsymbol{\theta})}{\pi_\rho(\boldsymbol{\theta})} \leq C_\rho$ for every $\boldsymbol{\theta} \in \mathbb{R}^d$, where

$$C_\rho := \exp\left(d\sigma_U^2 + \rho^4(2+d)\sigma_U^4\right) \cdot \prod_{i=1}^{b}(1 + \rho^2 M_i)^{d_i/2} \cdot \frac{\det\left(\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{1/2}}{\det\left(\sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{1/2}}, \tag{61}$$

with σ_U^2 defined as in (15).

Proof Let U^ρ be defined as in (7). By (33) and (31), we have

$$\begin{aligned}
\exp(-U^\rho(\boldsymbol{\theta})) &\leq \exp\left(-U(\boldsymbol{\theta})\right) \cdot \prod_{i=1}^{b} \frac{1}{\left(1 + \rho^2 m_i\right)^{d_i/2}} \cdot \exp\left(\sum_{i=1}^{b} \frac{\rho^2 \left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2}{2(1+\rho^2 m_i)}\right), \\
\exp(-U^\rho(\boldsymbol{\theta})) &\geq \exp\left(-U(\boldsymbol{\theta})\right) \cdot \prod_{i=1}^{b} \frac{1}{\left(1 + \rho^2 M_i\right)^{d_i/2}} \cdot \exp\left(\sum_{i=1}^{b} \frac{\rho^2 \left\|\nabla U_i(\mathbf{A}_i\boldsymbol{\theta})\right\|^2}{2(1+\rho^2 M_i)}\right).
\end{aligned} \tag{62}$$

Using (62), we have

$$\begin{aligned}
\pi_\rho(\boldsymbol{\theta}) &= \frac{\exp(-U^\rho(\boldsymbol{\theta}))}{Z_\rho} \\
&\geq \frac{\exp(-U(\boldsymbol{\theta}))}{Z_\rho} \cdot \frac{1}{\prod_{i=1}^b (1+\rho^2 M_i)^{d_i/2}} \\
&\geq \frac{\exp\left(-U(\boldsymbol{\theta}^\star) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top (\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i)(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)\right)}{Z_\rho} \cdot \frac{1}{\prod_{i=1}^b (1+\rho^2 M_i)^{d_i/2}}.
\end{aligned} \tag{63}$$

To lower bound $\pi_\rho(\boldsymbol{\theta})$, we need to upper bound Z_ρ. Using Lemma 21, we can do this based on an upper bound on Z. Using (A_4), we have

$$\begin{aligned}
Z &= \int_{\mathbb{R}^d} \exp\left(-\sum_{i=1}^{b} U_i(\mathbf{A}_i\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} \\
&\leq \exp\left(-\sum_{i=1}^{b} U_i(\mathbf{A}_i\boldsymbol{\theta}^\star)\right) \int_{\mathbb{R}^d} \exp\left(-\sum_{i=1}^{b} \frac{m_i}{2} \left\|\mathbf{A}_i\boldsymbol{\theta} - \mathbf{A}_i\boldsymbol{\theta}^\star\right\|^2\right) \mathrm{d}\boldsymbol{\theta} \\
&= \exp\left(-U(\boldsymbol{\theta}^\star)\right) (2\pi)^{d/2} \det\left(\sum_{i=1}^{b} m_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2}.
\end{aligned} \tag{64}$$

Note that the proposal density is of the form

$$\begin{aligned}\nu(\boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^\star, \left(\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1}\right) \\ &= \exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)^\top (\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i)(\boldsymbol{\theta}-\boldsymbol{\theta}^\star)\right) \frac{\det\left(\sum_{i=1}^b M_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{1/2}}{(2\pi)^{d/2}}. \end{aligned} \tag{65}$$

Under the assumption that $b = 1$, $d_1 = d$ and $\mathbf{A}_1$ is of full rank, we have $Z_\rho = Z$ by Lemma 21. The claim of the lemma in this single splitting case follows by comparing (65), (63) and using (64).

More generally, from Lemma 21, it follows for $\rho^2 \leq \frac{1}{6\sigma_U^2}$ that

$$\begin{aligned} Z_\rho &\leq Z \exp\left(\mathbb{E}_\pi(\bar{B}(\boldsymbol{\theta})) + \rho^4(2+d)\sigma_U^4\right) \\ &\leq \exp\left(-U(\boldsymbol{\theta}^\star)\right)(2\pi)^{d/2}\det\left(\sum_{i=1}^b m_i \mathbf{A}_i^\top \mathbf{A}_i\right)^{-1/2} \exp\left(d\sigma_U^2 + \rho^4(2+d)\sigma_U^4\right), \end{aligned}$$

where, in the last line, we used the fact that $\mathbb{E}_\pi(\bar{B}(\boldsymbol{\theta})) \leq d\sigma_U^2$, see Lemma 20. The claim of the lemma in this multiple splitting case now follows by comparing (65) and (63), and using the above upper bound on Z_ρ. ■

Now we are ready to prove our convergence bound in total variation distance.

Proof [Proof of Theorem 14] From Propositions 8 and 7, a sufficient condition to satisfy $\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq \epsilon/2$ is to have

$$\rho^2 \leq \frac{\epsilon}{dM_1}.$$

From Corollary 25, we know that the absolute spectral gap of SGS satisfies that $\gamma^* \geq K_{\mathrm{SGS}}$ (defined in (19)), and Proposition 26 implies that

$$\begin{aligned} \left\|\nu \mathbf{P}_{\mathrm{SGS}}^t - \pi_\rho\right\|_{\mathrm{TV}} &\leq \sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right] - 1} \cdot (1-\gamma^*)^t \\ &\leq \sqrt{\mathbb{E}_\nu\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)} \cdot (1-K_{\mathrm{SGS}})^t \\ &\leq \sqrt{C_\rho}(1-K_{\mathrm{SGS}})^t, \end{aligned}$$

where in the last step we have used Lemma 31 (C_ρ is defined as in Equation 60). By some algebra, using the definition of $t_{\mathrm{mix}}(\epsilon;\nu)$, and the fact that $\frac{1}{\log(1/(1-x))} \leq \frac{1}{x}$ for $0 < x < 1$, the above bound implies that

$$\left\|\nu \mathbf{P}_{\mathrm{SGS}}^{t(\epsilon)} - \pi_\rho\right\|_{\mathrm{TV}} \leq \frac{\epsilon}{2},$$

with the choice

$$t \geq \frac{\log\left(\frac{2}{\epsilon}\right) + C/2}{K_{\mathrm{SGS}}}. \tag{66}$$

Here

$$C = \frac{5d}{8} + \frac{d}{2}\log\left(\frac{M_1}{m_1}\right).$$

With the above choice for ρ^2 and the condition (66), the claim of Theorem 14 then follows by the triangle inequality. ■

Proof [Proof of Theorem 15] From (16), we have for $\rho^2 \leq \frac{1}{6\sigma_U^2}$,

$$\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq \rho^2 \frac{1}{2}\sum_{i=1}^{b} d_i M_i + \rho^4 \sigma_U^4 \left(2 + \frac{3}{2}d\right).$$

Then, a sufficient condition to satisfy $\|\pi_\rho - \pi\|_{\mathrm{TV}} \leq \epsilon/2$ is to have

$$\begin{aligned}
\rho^2 \frac{1}{2}\sum_{i=1}^{b} d_i M_i + \rho^4 \sigma_U^4 \left(2 + \frac{3}{2}d\right) &\leq \frac{\epsilon}{2} \\
\rho^4 \sigma_U^4 \left(2 + \frac{3}{2}d\right) + \frac{1}{2}\rho^2 \sum_{i=1}^{b} d_i M_i - \frac{\epsilon}{2} &\leq 0 \\
R^2 \sigma_U^4 \left(2 + \frac{3}{2}d\right) + R\frac{1}{2}\sum_{i=1}^{b} d_i M_i - \frac{\epsilon}{2} &\leq 0, \quad \text{with } R = \rho^2.
\end{aligned}$$

This inequality is satisfied under the condition

$$\rho^2 \leq \frac{\sum_{i=1}^{b} d_i M_i \left(\sqrt{1 + 8\epsilon\sigma_U^4\left(2 + \frac{3}{2}d\right)\left(\sum_{i=1}^{b} d_i M_i\right)^{-2}} - 1\right)}{4\sigma_U^4\left(2 + \frac{3}{2}d\right)} \wedge \frac{1}{6\sigma_U^2}.$$

From Corollary 25, we know that the absolute spectral gap of SGS satisfies that $\gamma^* \geq K_{\mathrm{SGS}}$ (defined in (19)), and Proposition 26 implies that

$$\begin{aligned}
\left\|\nu \mathbf{P}_{\mathrm{SGS}}^t - \pi_\rho\right\|_{\mathrm{TV}} &\leq \sqrt{\mathbb{E}_{\pi_\rho}\left[\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)^2\right] - 1} \cdot (1 - \gamma^*)^t \\
&\leq \sqrt{\mathbb{E}_{\nu}\left(\frac{\mathrm{d}\nu}{\mathrm{d}\pi_\rho}\right)} \cdot (1 - K_{\mathrm{SGS}})^t \\
&\leq \sqrt{C_\rho}(1 - K_{\mathrm{SGS}})^t,
\end{aligned}$$

where in the last step we have used Lemma 31 (C_ρ is defined as in (61)). Again, by some algebra, using the definition of $t_{\text{mix}}(\epsilon;\nu)$, and the fact that $\frac{1}{\log(1/(1-x))} \leq \frac{1}{x}$ for $0 < x < 1$, the above bound implies that

$$\left\| \nu \mathbf{P}_{\text{SGS}}^{t(\epsilon)} - \pi_\rho \right\|_{\text{TV}} \leq \frac{\epsilon}{2},$$

with the choice

$$t \geq \frac{\log\left(\frac{2}{\epsilon}\right) + C/2}{K_{\text{SGS}}}. \tag{67}$$

Here

$$C = d\sigma_U^2 + \rho^4(2+d)\sigma_U^4 + \frac{17}{32}\sum_{i=1}^{b} d_i + \frac{1}{2}\log\left(\frac{\det\left(\sum_{i=1}^{b} M_i \mathbf{A}_i^\top \mathbf{A}_i\right)}{\det\left(\sum_{i=1}^{b} m_i \mathbf{A}_i^\top \mathbf{A}_i\right)}\right).$$

With the above choice for ρ^2 and the condition (67), the claim of Theorem 15 then follows by the triangle inequality. ■

Appendix C.6. Additional Details for the Toy Gaussian Example

This section gives additional details concerning the results depicted on Figure 2. For each splitting strategy introduced in Section 3.2, we give explicit formulas for the bounds on both TV and 1-Wasserstein distances.

Appendix C.6.1. Splitting Strategy 1

Starting from an initial value $\theta^{[0]} \sim \nu$, we now show the explicit form of the Markov transition kernel νP_{SGS}^t after t iterations. To this purpose, we take advantage that the θ-chain corresponds in this case to an auto-regressive process of order 1. Indeed, the conditional distributions of θ and $\mathbf{z}_{1:b}$ writing

$$\Pi_\rho(\mathbf{z}_i|\theta) = \mathcal{N}\left(\mathbf{z}_i; \frac{\mu\rho^2 + \theta\sigma^2}{\sigma^2+\rho^2}, \frac{\rho^2\sigma^2}{\rho^2+\sigma^2}\right), \forall i \in [b]$$

$$\Pi_\rho(\theta|\mathbf{z}_{1:b}) = \mathcal{N}\left(\theta; \bar{z}, \frac{\rho^2}{b}\right), \text{ where } \bar{z} \coloneqq \frac{1}{b}\sum_{i=1}^{b}\mathbf{z}_i,$$

we have

$$P_{\text{SGS}} \coloneqq \Pr\left(\theta^{[t]}|\theta^{[t-1]}\right) = \mathcal{N}\left(\theta^{[t]}; \frac{\sigma^2}{\sigma^2+\rho^2}\theta^{[t-1]} + \frac{\rho^2}{\sigma^2+\rho^2}\mu, \frac{2\rho^2\sigma^2+\rho^4}{b(\rho^2+\sigma^2)}\right).$$

By a straightforward induction, it follows that the Markov transition kernel νP^t after t iterations and with initial distribution ν has the form

$$\begin{aligned}
&\nu P_{\text{SGS}}^t \coloneqq \Pr\left(\theta^{[t]}|\theta^{[0]} \sim \nu\right) \\
&= \mathcal{N}\left(\theta^{[t]}; \left(\frac{\sigma^2}{\sigma^2+\rho^2}\right)^t \theta^{[0]} + \frac{\rho^2\mu}{\sigma^2+\rho^2}\sum_{i=0}^{t-1}\left(\frac{\sigma^2}{\sigma^2+\rho^2}\right)^i, \frac{2\rho^2\sigma^2+\rho^4}{b(\rho^2+\sigma^2)}\sum_{i=0}^{t-1}\left(\frac{\sigma^4}{(\sigma^2+\rho^2)^2}\right)^i\right).
\end{aligned}$$

Appendix C.6.2. Splitting Strategy 2

Similar calculus as in the above section can be undertaken by simply replacing ρ^2 by $\rho^2 b$.

References

Manya V. Afonso, José M. Bioucas-Dias, and Mário A. T. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(9):2345–2356, 2010.

Jim H. Albert and Siddartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Ludwig Arnold. *Stochastic Differential Equations: Theory and Applications*. Wiley-Interscience, New York, 1974.

Daniel Azagra, Juan Ferrera, Fernando López-Mesas, and Yenny Rangel. Smooth approximation of Lipschitz functions on Riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 326(2):1370–1378, 2007.

Jack Baker, Paul Fearnhead, Emily B. Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615, 2019.

Andrei-Cristian Barbos, François Caron, Jean-François Giovannelli, and Arnaud Doucet. Clone MCMC: parallel high-dimensional Gaussian Gibbs sampling. In *Advances in Neural Information Processing Systems 30*, pages 5020–5028, 2017.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25:185–209, 2015.

Jonathan Bennett and Neal Bez. Generating monotone quantities for the heat equation. *Journal für die Reine und Angewandte Mathematik*, 2015.

Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B*, 48(3):259–279, 1986.

Joris Bierkens, Paul Fearnhead, and Gareth Roberts. The Zig-Zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics*, 47:1288–1320, 2019.

Alexandre Bouchard-Côté, Sebastian J. Vollmer, and Arnaud Doucet. The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 113:855–867, 2018.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems 31*, pages 8268–8278, 2018.

Niladri S. Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L. Bartlett, and Michael I. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Zongchen Chen and Santosh S. Vempala. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. In *Proceedings of the International Conference on Randomization and Computation*, 2019.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 186–211, 2018.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 300–323, 2018.

Hee Min Choi and Jim Hobert. The Pólya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

Edmond Chow and Yousef Saad. Preconditioned Krylov subspace methods for sampling multivariate Gaussian distributions. *SIAM Journal on Scientific Computing*, 36(2):A588–A608, 2014.

Arkabandhu Chowdhury and Christopher Jermaine. Parallel and distributed MCMC via shepherding distributions. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1819–1827, 2018.

Robert Cornish, Paul Vanetti, Alexandre Bouchard-Côté, George Deligiannidis, and Arnaud Doucet. Scalable Metropolis Hastings for exact Bayesian inference with large datasets. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1351–1360, 2019.

Paolo Dai Pra, Benedetto Scoppola, and Elisabetta Scoppola. Sampling from a Gibbs measure with pair interaction by means of PCA. *Journal of Statistical Physics*, 149(4): 722–737, 2012.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society, Series B*, 79(3):651–676, 2017.

Arnak S. Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and Their Applications*, 129(12): 5278–5311, 2019.

Arnak S. Dalalyan and Lionel Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Kumar Avinava Dubey, Sashank J. Reddi, Sinead A. Williamson, Barnabas Poczos, Alexander J. Smola, and Eric P. Xing. Variance reduction in stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems 29*, pages 1154–1162, 2016.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Alain Durmus, Szymon Majewski, and Blażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J. Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20(183): 1–42, 2019.

Wally R. Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society, Series C*, 41(2):337–348, 1992.

Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2862–2869, 2014.

Timothy E. Hanson, Adam J. Branscum, and Wesley O. Johnson. Informative g -priors for logistic regression. *Bayesian Analysis*, 9(3):597–612, 09 2014.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18(1):3744–3780, 2017.

Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 03 2006.

De Huang and Joel A. Tropp. Nonlinear matrix concentration via semigroup methods. *Electronic Journal of Probability*, 26:1 – 31, 2021.

Milos Ilic, Tony Pettitt, and Ian Turner. Bayesian computations and efficient algorithms for computing functions of large sparse matrices. *ANZIAM Journal*, 45(E):504–518, 2004.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Gregory F. Lawler. *Random Walk and the Heat Equation*. American Mathematical Society, 2010.

Michel Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, 2001.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Proceedings of 34th Conference on Learning Theory*, pages 2993–3050, 2021.

Claude Lemaréchal and Claudia Sagastizábal. Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

Paul Lévy. Sur certains processus stochastiques homogènes. *Compositio Mathematica*, 7: 283–339, 1940.

Qing Li and Nan Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. *arXiv preprint arXiv:2110.04597*, 2021.

Jun S. Liu, Wing Hung Wong, and Augustine Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.

Chris J. Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

Luca Martino and Joaquín Míguez. A generalization of the adaptive rejection sampling algorithm. *Statistics and Computing*, 21:633–647, 2011.

Peter McCullagh and John A Nelder. *Generalized Linear Models*. Routledge, 2019.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Willie Neiswanger, Chong Wang, and Eric P. Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

A. Parker and C. Fox. Sampling Gaussian distributions in Krylov spaces with conjugate gradients. *SIAM Journal on Scientific Computing*, 34(3):B312–B334, 2012.

Grigorios A. Pavliotis. *Diffusion Processes, the Fokker-Planck and Langevin Equations*. Springer, New York, 2014.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

Lawrence Perko. *Differential Equations and Dynamical Systems*. Springer Science & Business Media, 2013.

Vincent Plassier, Maxime Vono, Alain Durmus, and Eric Moulines. DG-LMC: a turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8577–8587, 2021.

Nicholas Polson, James Scott, and Jesse Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526): 831–843, 2019.

Lewis Rendell, Adam Johansen, Anthony Lee, and Nick Whiteley. Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, 30(2):249–259, 2021.

Herbert Robbins. A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29, 1955.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

Gareth O. Roberts and Adrian F.M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and Their Applications*, 49(2):207–216, 1994.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.

Ralph Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer-Verlag, Berlin, 1998.

Daniel Sabanes Bove and Leonhard Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410, 09 2011.

H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947.

Steven L. Scott. Comparing consensus Monte Carlo strategies for distributed Bayesian computation. *Brazilian Journal of Probability and Statistics*, 31(4):668–685, 2017.

Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016.

Leonida Tonelli. Sull'integrazione per parti. *Rendiconti. Accademia Nazionale dei Lincei*, 5(18):246–253, 1909.

Tristan van Leeuwen and Felix J. Herrmann. A penalty method for PDE-constrained optimization in inverse problems. *Inverse Problems*, 32(1):015007, 2015.

Luis Vargas, Marcelo Pereyra, and Konstantinos C. Zygalakis. Accelerating proximal Markov chain Monte Carlo by using explicit stabilised methods. *SIAM Journal on Imaging Sciences*, 13(2):905–935, 2020.

Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Split-and-augmented Gibbs sampler–Application to large-scale inference problems. *IEEE Transactions on Signal Processing*, 67(6):1648–1661, 2019.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Asymptotically exact data augmentation: models, properties and algorithms. *Journal of Computational and Graphical Statistics*, 30(2):335–348, 2021a.

Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. High-dimensional Gaussian sampling: a review and a unifying approach based on a stochastic proximal point algorithm. *SIAM Review*, 2021b. Forthcoming.

Xiangyu Wang and David B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Xiangyu Wang, Fangjian Guo, Katherine A. Heller, and David B. Dunson. Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems 28*, 2015.

Yilun Wang, Junfeng Yang, Wotao Yin, and Yin Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

Xiaofan Xu and Malay Ghosh. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015.